Abstract

The theory and methods of digital signal processing (DSP) are becoming increasingly important in molecular biology. However, since DNA sequences are strings of characters, numerical values should be associated with the sequences before techniques in DSP can be applied to DNA sequence analysis. Ways of conversion varies, but each with its suitable applications. Intended for engineers devoted to signal processing, some related fundamentals in molecular biology are presented first in this tutorial, methods of associating numerical sequences to DNA sequences are introduced next, followed by typical topics studied via these methods. Some concluding remarks are addressed in the end.

Shang-Ching Lin

Graduate Institute of Biomedical Electronics and Bioinformatics,
National Taiwan University, Taipei, Taiwan
Contents

1 Introduction 2

2 Some Biological Fundamentals 3
  2.1 DNA 3
  2.2 The Central Dogma 5
  2.3 Material: Public Databases 7

3 From Character Strings To Numerical Values 8
  3.0 DNA Sequence and DSP: Protein Coding as an Example 8
  3.1 Indicator Sequence 10
  3.2 Symbolic Autocorrelation 10
  3.3 Spectral Envelope 11
  3.4 Issue: Reduction of the Dimensionality 12
  3.5 Electron-Ion Interacting Potential (EIIP) 13
  3.6 DNA Walk 14
  3.7 Huffman coding 16

4 Typical Topics 17
  4.1 Gene Prediction: Identification of Protein Coding DNA Regions 17
    4.1.1 Characteristics of protein coding DNA regions 17
    4.1.2 DNA Filtering Example: IIR Antinotch Filter 17
    4.1.3 DNA spectrogram 18
    4.1.4 Identification of Protein Coding DNA Regions 20
  4.2 Identification of Reading Frame 22
    4.2.1 Color coding and color map approach 22
  4.3 Prediction of Gene Function 24
  4.4 Long-range correlation 24
  4.5 Study on Gene Regulation 25

5 Concluding Remarks 26

6 Reference 27

Glossary 28

1 The picture on the cover is taken from reference [b].
1 Introduction

Genomics\textsuperscript{2} is a highly cross-disciplinary field that creates paradigm shifts in such diverse areas as medicine and agriculture. It is believed that many significant scientific and technological endeavors in the 21\textsuperscript{st} century will be related to the processing and interpretation of the vast information that is currently revealed from sequencing the genomes of many living organisms, including humans. Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been mathematically represented by character strings, in which each character is a letter of an alphabet. In the case of DNA, the alphabet is size 4 and consists of the letters A, T, C and G; in the case of proteins, the size of the corresponding alphabet is 20.

Biomolecular sequence analysis has already been a major research topic among computer scientists, physicists, and mathematicians. The main reason that the field of signal processing does not yet have significant impact in the field is because it deals with numerical sequences rather than character strings. However, if we properly map a character string into one or more numerical sequences, then digital signal processing (DSP) provides a set of novel and useful tools for solving highly relevant problems.

For example, in the form of local texture, color spectrograms visually provide significant information about biomolecular sequences which facilitates understanding of local nature, structure, and function. Furthermore, both the magnitude and the phase of properly defined Fourier transforms can be used to predict important features like the location and certain properties of protein coding regions in DNA. Even the process of mapping DNA into proteins and the interdependence of the two kinds of sequences can be analyzed using simulations based on digital filtering. These and other DSP-based approaches result in alternative mathematical formulations and may provide improved computational techniques for the solution of useful problems in genomic information science and technology.

This tutorial is intended for engineers devoted to signal processing, thus some related fundamentals in molecular biology are presented first. As DNA sequences are character strings, for DSP techniques to be applicable to these data, methods converting DNA sequences to numerical sequences are required and will be presented next. Typical research topics utilizing these methods are summarized in section 4, and the tutorial concludes with some remarks.

\textsuperscript{2} For meanings of terms in molecular biology or bioinformatics, please refer to the glossary provided at the end of this tutorial. Words in the glossary will be underlined upon their first appearance in the text.
2 Some Biological Fundamentals

In this section, basic ideas regarding DNA (deoxyribonucleic acid) are introduced. DNA molecules store the digital information that constitutes the genetic blueprint of living organisms, and they are “realized” through a process described by the famous “Central Dogma” in molecular biology, which is introduced subsequently. A vast amount of DNA/protein sequence and protein 3-D structure data are available on the Internet as online databases and are free of charge, accessible to any interested individual. This promotes the study in this area and a short introduction of it is given in the last part of this section.

2.1 DNA

A single strand of DNA is a biomolecule consisting of many linked, smaller components called nucleotides. Each nucleotide is one of four possible types designated by the letters A, T, C, and G and has two distinct ends, the 5′ end and the 3′ end, so that the 5′ end of a nucleotide is linked to the 3′ end of another nucleotide by a strong chemical bond (covalent bond), thus forming a long, one-dimensional chain (backbone) of a specific directionality. Therefore, each DNA single strand is mathematically represented by a character string, which, by convention specifies the 5′ to 3′ direction when read from left to right.

Single DNA strands tend to form double helices with other single DNA strands. Thus, a DNA double strand contains two single strands called complementary to each other because each nucleotide of one strand is linked to a nucleotide of the other strand by a chemical bond (hydrogen bond), so that A is linked to T and vice versa, and C is linked to G and vice versa. Each such bond is weak (contrary to the bonds forming the backbone), but together all these bonds create a stable, double helical structure. The two strands run in opposite directions, as shown in Fig. 1, in which we see the sugar-phosphate chemical structure of the DNA backbone created by strong (covalent) bonds, and that each nucleotide is characterized by a base that is attached to it. The two strands are linked by a set of weak (hydrogen) bonds. The bottom left diagram is a simplified, straightened out depiction of the two linked strands.

For example, the part of the DNA double strand shown in Fig. 1 is

\[
\begin{align*}
5′ & \quad C-A-T-T-G-C-C-A-G-T \quad 3′ \\
3′ & \quad G-T-A-A-C-G-G-T-C-A \quad 5′
\end{align*}
\]

Because each of the strands of a DNA double strand uniquely determines the other strand, a double-stranded DNA molecule is represented by either of the two character strings read in its 5′ to 3′ direction. Thus, in the example above, the character strings

\footnote{Refer the term “DNA” in the glossary for some more specific information.}
CATTGCCAGT and ACTGGCAATG can be alternatively used to describe the same DNA double strand, but they specify two different single strands which are complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or **palindromes**. For example AATCTAGATT is a palindrome.

DNA molecules store the **digital information** that constitutes the genetic blueprint of living organisms. This digital information has been created and reliably stored throughout billions of years of evolution during which some vital regions of DNA sequences have been remarkably preserved, despite striking differences in the body plans of various animals.

A DNA sequence can be separated into two types of regions: **genes** and intergenic spaces. Genes contain the information for generation of proteins. Each gene is responsible for the production of a different protein. Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells. For example the set of genes that are active in blood cells are different from those that are active in nerve cells, which explains why these cells look so different!
2.2 The Central Dogma

Fig. 2 shows some of the steps involved in the production of a protein from a gene. Notice that a gene has two types of subregions called the exons and introns (procaryotes like bacteria do not have introns). The gene is first transcribed into a single stranded chain called the messenger RNA or mRNA molecule. The introns are then removed from the mRNA by a process called splicing.

The spliced mRNA is divided into groups of three adjacent bases. Each triplet is called a codon. Evidently there are 64 possible codons. Thus the mRNA is nothing but a sequence of codons. Each codon instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. This mapping is called the genetic code and is shown in Fig. 3. Since there are 64 possible codons but only 20 amino acids, the mapping from codons to amino acids is many-to-one. Notice that there is a start codon ATG which signifies the beginning of the protein-coding part of the gene. If a start codon occurs inside a gene again, it produces the amino acid methionine. A stop codon signifies that the protein coding part of the gene has come to an end. There are three stop codons. It is a wonder of Nature that all life forms (from bacteria to mammals) use the same genetic code. This is no doubt due to the common origin of all life.

The translation from mRNA to protein is aided by adaptor molecules called the transfer RNA or tRNA molecules. In some sense the tRNA molecules store the genetic code. One end of the tRNA matches a specific codon and the other end attaches to the corresponding amino acid. See Fig. 4. The molecule ribosome works in conjunction with tRNA molecules and mRNA to produce the protein. So it is clear that the genetic code is essentially stored in the tRNA molecules. And due to their function, ribosomes are often referred to as the protein factories of the cell. There are many ribosomes in a cell working in parallel like molecular machines.

Many details are omitted in Fig. 2 for brevity. For example the mRNA is in reality the complement of the gene, that is, Cs are replaced with Gs, and As with Ts (rather Us). Thus, if the gene is ATTAGC then the mRNA is UAAUCG. There is a second level of complementing which cancels this when the mRNA attaches to tRNA molecules at the so-called anticodon sites.

The observation that each gene is responsible for the creation of a protein (through mRNA) is often expressed as

\[
\begin{align*}
\text{gene in DNA} & \quad \overset{\text{transcription}}{\rightarrow} \quad \text{RNA} & \quad \overset{\text{translation}}{\rightarrow} \quad \text{protein}
\end{align*}
\]

and is referred to as the central dogma of molecular biology.
Fig. 2  When a gene is ready to be expressed, it is duplicated in the form of a single strand molecule called the mRNA (messenger RNA) which then leaves the nucleus. The introns are spliced out and a shorter mRNA molecule is produced. Thus, unlike the parent gene, the mRNA is a concatenation of the exons only. It is used by ribosomes outside the nucleus of the cell to manufacture the appropriate protein coded by the original gene. Thus protein production involves the transcription of genes into mRNA and the subsequent translation of the 4-letter language to a 20-letter language. [2]

Fig. 3  The genetic code. Triples of bases such as AAA denote codons. The single letters such as K denote amino acids. Their three letter names (e.g., Lys) are also shown [2]. Full names of amino acids can be found in Figure 8. of [2].
Fig. 4  Example of a transfer RNA molecule in yeast. The bases are numbered from 1 to 76. Only a particular codon can match perfectly with the anticodon, and can therefore be associated with the specific amino acid that is able to attach to the tRNA at the top end. In this manner, the tRNA molecules store the genetic code in the cell. [2]

2.3  Material: Public Databases [1]

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH). The NIH nucleotide sequence database is called GenBank and contains all publicly available DNA sequences. For example, one can go to http://www.ncbi.nlm.nih.gov/entrez and identify the DNA sequence with Accession Number AF 099922; choose Nucleotide under Search and then fill out the other entry by typing: AF 099922 [Accession] and pressing “Go.” Clicking on the resulting accession number will show the annotation for the genes as well as the whole nucleotide sequence in the form of raw data. Similarly, Entrez provides access to databases of protein sequences as well as 3-D macromolecular structures, among other options. As another example, a specialized repository for the processing and distribution of 3-D, macromolecular structures can be found in the Protein Data Bank at www.rcsb.org.
3 From Character Strings To Numerical Values

3.0 DNA Sequence and DSP: Protein Coding as an Example \[1\]

In a DNA sequence of length N, assume that we assign the numbers a, t, c, g to the characters A, T, C, G, respectively. A proper choice of the numbers a, t, c and g can provide potentially useful properties to the numerical sequence $x[n]$.

For example, if we choose complex conjugate pairs $t=a^*$ and $g=c^*$, then the complementary DNA strand is represented by

$$\tilde{x}[n] = x^*[-n + N - 1], \quad n = 0, 1, \ldots, N - 1$$

and, in this case, all palindromes will yield conjugate, symmetric numerical sequences which have interesting mathematical properties, including generalized linear phase.

One such assignment (the simplest out of many possible ones) is the following:

$$a = 1 + j, \quad t = 1 - j, \quad c = -1 - j, \quad g = -1 + j.$$ \hspace{1cm} (2)

We may also assign numerical values to amino acids by modeling the protein coding process as an FIR digital filter, in which the input $x[n]$ is the numerical nucleotide sequence, and the output $y[n]$ is the possible resulting numerical amino acid sequence (if $x[n]$ is within a coding region in the proper reading frame):


For example, if we set $h[0]=1$, $h[1]=1/2$, and $h[2]=1/4$, and $x[n]$ is defined by the parameters in (2), then $y[n]$ can only take one out of 64 possible values.

Furthermore, if for example, $x[n]$ corresponds to a forward coding DNA sequence in the first reading frame (i.e., if $x[0]$, $x[1]$, $x[2]$ corresponds to the first codon), then the elements of the output subsequence: $y[2]$, $y[5]$, $y[8]$, $y[11]$, ..., $y[N-1]$ are complex numbers representing each of the amino acids of the resulting protein. In fact, the entire genetic code can be drawn on the complex plane as shown in Fig. 5, in which the center of the square labeled Met (coded by ATG), is the complex number $(1+j) + 0.5(1-j) + 0.25(-1+j) = 1.17+0.88j$.

Each of the entries in Fig. 5 correspond to one of the 20 amino acids or the STOP codon. Therefore, the protein coding process can be simulated by a digital low-pass filter, followed by subsampling via a three-band polyphase decomposition, followed by a switch selecting one of the three bands (reading frames), followed by a vector quantizer as defined in Fig. 5.

The simplest way of performing Fourier or any other transform analysis on a symbolic sequence is to map the symbols to numbers, and then process the sequence.
obtained. For example, one could start by finding the autocorrelation of the numeric sequence, and its Fourier transform. This has disadvantages – the mapping may either expose or hide some of the frequency information\(^4\). Furthermore, there might be no **biochemical meaning** for the ordering and arithmetic structure that result from the symbolic to numeric mapping.

![The genetic code on the complex plane. [1]](image)

In sections 3.1 to 3.4, the following notations are used:

The Discrete Fourier Transform (DFT) \(X[k], 0 \leq k < n\) of the input sequence \(x[k], 0 \leq k < n\) is denoted by \(X = FX\), where \(F\) is the \(n \times n\) Fourier matrix defined as

\[
F_{ab} = e^{-j\frac{2\pi ab}{n}}, \quad a, b = 0, 1, \ldots, n - 1.
\]

On the other hand, the inner product is denoted by

\[
\langle x, y \rangle = \sum_{k=0}^{n-1} \bar{x}_k y_k = x' y
\]

where the upper bar denotes conjugation and the prime denotes conjugate transpose.

\(^4\) For an example illustrating this, see section 1 in [3].
Thus we have

\[ \langle x, x \rangle = \|x\|^2 = \frac{1}{n} \|X\|^2 \]

### 3.1 Indicator Sequence [3]

The indicator sequence of adenine (A) of a DNA sequence \( s_k, 0 \leq k < n \) is defined as

\[
\nu_k^A = \begin{cases} 
1, & s_k = A, \\
0, & s_k \neq A 
\end{cases}
\]

and similarly for the other three bases.

The total spectrum of a symbolic sequence is often defined as the squared modulus of the DFT’s of the indicator sequences, that is

\[
R_i = |U_i^A|^2 + |U_i^G|^2 + |U_i^C|^2 + |U_i^T|^2
\]

with \( i \in \mathbb{Z}_N \). In the literature, the spectrum is sometimes identified, with little or no explanation, with this sum. Intuitively, the solution seems reasonable. No algebraic operations need to be defined on the symbols, and no symbolic-to-numeric mapping is needed. However, the theoretical interpretation and meaning of this solution seems, at first glance, obscure.

### 3.2 Symbolic Autocorrelation [3]

The symbolic autocorrelation method allows a more satisfactory view of the indicator sequence approach. As we will see, under that formulation the spectrum (4) emerges as the Fourier transform of the symbolic autocorrelation.

By autocorrelation of the symbolic sequence \( s_k, 0 \leq k < n \) we mean the numeric sequence

\[
r_k = \sum_{i=0}^{n-1} d(s_i, s_{i+k})
\]

where for any two symbols \( x \) and \( y \)

\[
d(x, y) = \begin{cases} 
1, & x = y, \\
0, & x \neq y
\end{cases}
\]

The autocorrelation \( r_k \) is a numeric sequence that can be rewritten in terms of the four indicator sequences, since \( u_k^A = d(s_k, A) \), \( u_k^G = d(s_k, G) \), and so on. Hence,

\[
r_k = \langle u^A, S_k u^A \rangle + \langle u^C, S_k u^C \rangle + \langle u^G, S_k u^G \rangle + \langle u^T, S_k u^T \rangle
\]
where the operator $S_k$ denotes a cyclic shift by $k$, that is, $S_k u^a = (u_{m+k}^a)_{0 \leq m < n}$, and similarly in the remaining three cases. By the definition of inner product, we have

$$\langle u^a, S_k u^a \rangle = \frac{1}{n} \sum_{i=0}^{n-1} |U_i^a|^2 e^{-j \frac{2\pi}{n} ki}$$

and so

$$r_k = \frac{1}{n} \sum_{i=0}^{n-1} (|U_i^a|^2 + |U_i^c|^2 + |U_i^g|^2 + |U_i^f|^2) e^{-j \frac{2\pi}{n} ki}$$

We conclude that the DFT of the symbolic autocorrelation is the sum of the squared modulus of the DFT’s of the indicator sequences. In other words, we obtain (4), the total spectrum using the indicator sequences.

### 3.3 Spectral Envelope [3]

Consider the $n \times 4$ matrix

$$u = [u^a \ u^c \ u^g \ u^f]$$

and the vector of real weights

$$w = [a \ c \ g \ t]^T.$$  

The sequence $z = uw$ then corresponds to the mapping $A \mapsto a$, $C \mapsto c$, and so on. The DFT of $z$ is

$$Z = Fz = Fuw = Uw,$$

where $U$ is the $4 \times n$ matrix obtained by concatenating the DFTs of the indicator sequences,

$$U = Fu = [Fu^a \ Fu^c \ Fu^g \ Fu^f] = [U^a \ U^c \ U^g \ U^f].$$

Denoting by $U_i$ the $i$th line of $U$, we may write $Z_i = U_iw$, and so

$$|Z_i|^2 = w' U_i' U_i w = |a U_i^a + c U_i^c + g U_i^g + t U_i^f|^2$$

The idea underlying the study in 4.1 and 4.2 is to adjust the symbolic-to-numeric mapping in such a way that the $Z_i$ become in some sense extremal. For each frequency $i$, select the vector $w$ of unit norm that maximizes $|Z_i|^2$. That is, consider the problem

$$\max_{\|w\|=1} |Z_i|^2 = \max_{\|w\|=1} w' U_i' U_i w$$
The maximum of this Rayleigh quotient is $\lambda_{\text{max}}(U_i'U_i)$, the maximum eigenvalue of the Hermitian matrix $U_i'U_i$. Furthermore, the weights $w$ for which the maximum is achieved are given by

$$w = \frac{U_i'}{\|U_i\|}$$

As a result,

$$\max_{\|w\|=1} |U_i'w|^2 = \max_{\|w\|=1} w'U_i'U_i w = |U_i'^a|^2 + |U_i'^c|^2 + |U_i'^g|^2 + |U_i'^t|^2 = R_i$$

and so we obtain

$$\lambda_{\text{max}}(U_i'U_i) = |U_i'^a|^2 + |U_i'^c|^2 + |U_i'^g|^2 + |U_i'^t|^2 = R_i$$

This reveals yet another way of looking at the total spectrum (4). We have seen that the sum of the squares of the DFTs of the four indicator sequences, at frequency $i$, is equal to the DFT of the symbolic autocorrelation, at frequency $i$. Now we see that it is also related to the value of the DFT of a certain numerical sequence, again at frequency $i$. The particular numerical sequence that leads to this spectrum corresponds to a symbolic-to-numeric mapping optimized to achieve the maximum squared magnitude for frequency $i$. This approach acts as the base of other more complicated approaches (for discussion on those, see references [8] and [10] in [3]), and is not strictly necessary here. To see this, apply the Cauchy inequality to (5),

$$|Z_i|^2 = |au_i'^a + cu_i'^c + gu_i'^g + iu_i'^t|^2$$

$$\leq (|a|^2 + |c|^2 + |g|^2 + |i|^2)(|U_i'^a|^2 + |U_i'^c|^2 + |U_i'^g|^2 + |U_i'^t|^2)$$

and then note that the condition for equality readily leads to the results.

### 3.4 Issue: Reduction of the Dimensionality [1, 3]

The four indicator sequences are of course redundant, since

$$u^a + u^c + u^g + u^t = 1$$

and so

$$U_i'^a + U_i'^c + U_i'^g + U_i'^t = \begin{cases} N, & i = 0, \\ 0, & i \neq 0. \end{cases}$$

The total spectrum can therefore be obtained with three DFT’s, rather than four. In fact, it is possible to work with three ($x$, $y$, $z$) nonredundant sequences, rather than with four redundant ones. The assignments used in [1] are
\[
\mathcal{A} \mapsto (0 \ 0 \ 1), \quad \mathcal{C} \mapsto \left( -\frac{\sqrt{2}}{3} \ -\frac{\sqrt{6}}{3} \ -\frac{1}{3} \right), \quad \mathcal{G} \mapsto \left( -\frac{\sqrt{2}}{3} \ -\frac{\sqrt{6}}{3} \ -\frac{1}{3} \right),
\]
\[
T \mapsto \left( \frac{2\sqrt{2}}{3} \ 0 \ -\frac{1}{3} \right),
\]

which are four vertices of a tetrahedron.

The connection with the indicator sequences is

\[
x = \frac{\sqrt{2}}{3}(2u^r - u^c - u^g), \quad y = \frac{\sqrt{6}}{3}(u^c - u^g), \quad z = \frac{1}{3}(3u^a - u^r - u^c - u^g). \tag{6}
\]

The equivalence between the method that relies on the indicator sequences, and those of reduced dimensionality was shown\(^5\), for an arbitrary number of symbols. In the present case, we have

\[
3(|U_i^a|^2 + |U_i^c|^2 + |U_i^g|^2) = \begin{cases} 
4(|X_i|^2 + |Y_i|^2 + |Z_i|^2), & i \neq 0, \\
4(|X_i|^2 + |Y_i|^2 + |Z_i|^2) - 1, & i = 0.
\end{cases}
\]

### 3.5 Electron-Ion Interacting Potential (EIIP) [4, 5]

The authors of [5] proposed a novel coding measure scheme by replacing the four binary indicator sequences by just one sequence which they call as “EIIP indicator sequence”.

The energy of delocalized electrons in amino acids and nucleotides has been calculated as the Electron-ion interaction pseudopotential (EIIP). The EIIP values of amino acids have already been used in Resonant Recognition Models (RRM) to substitute for the corresponding amino acids in protein sequences, whose Discrete Fourier Transforms are taken to extract the information contents. The Fourier cross spectra of a group of related proteins reveal a sharp peak at a frequency which is termed as the “characteristic frequency” of that group of proteins as they are found to represent a particular biological function and selectively interact with targets of the corresponding “characteristic frequency” (resonant recognition). This has been used to identify “hot spots” in proteins and for peptide design which are very useful in drug discovery. The EIIP values for the nucleotides are given in Table 1.

If we substitute the EIIP values for A, G, C & T in a DNA string \(x[n]\), we get a numerical sequence which represents the distribution of the free electrons’ energies along the DNA sequence. This sequence is named as the “EIIP indicator sequence”.

---

\(x_{c[n]}\). For example, if \(x[n] = A \ A \ T \ G \ C \ A \ T \ C \ A\), then using the values from Table 1, \(x_{c[n]} = [0.1260 \ 0.1260 \ 0.1335 \ 0.0806 \ 0.1340 \ 0.1260 \ 0.1335 \ 0.1340 \ 0.1260].\)

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>EIIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1260</td>
</tr>
<tr>
<td>G</td>
<td>0.0806</td>
</tr>
<tr>
<td>C</td>
<td>0.1340</td>
</tr>
<tr>
<td>T</td>
<td>0.1335</td>
</tr>
</tbody>
</table>

Table 1  Electron Ion Interaction pseudo potentials of nucleotides [5]

### 3.6 DNA Walk [6]

In order to study the scale-invariant long-range correlations of the DNA sequences, a graphical representation of DNA sequences is introduced in this paragraph, which was termed a “fractal landscape” or “DNA walk”. For the conventional one-dimensional random walk model, a walker moves either up \((u(i) = +1)\) or down \((u(i) = -1)\) one unit length for each step \(i\) of the walk. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each step depends on the history (“memory”) of the walker. The DNA walk is defined by the rule that the walker steps up \((u(i) = +1)\) if a pyrimidine\(^6\) occurs at position a linear distance \(i\) along the DNA chain, while the walker steps down \((u(i) = -1)\) if a purine\(^7\) occurs at position \(i\). The question we asked was whether such a walk displays only short-range correlations (as in an n-step Markov chain) or long-range correlations (as in critical phenomena and other scale-free “fractal” phenomena).

The DNA walk provides a graphical representation for each gene and permits the degree of correlation in the base pair sequence to be directly visualized, as in Fig. 6. This naturally motivates a quantification of this correlation by calculating the “net displacement” of the walker after \(l\) steps, which is the sum of the unit steps \(u(i)\) for each step \(i\). Thus,

\[
y(l) = \sum_{i=1}^{l} u(i).
\]

An important statistical quantity characterizing any walk is the root mean square fluctuation \(F(l)\) about the average of the displacement; \(F(l)\) is defined in terms of the difference between the average of the square and the square of the average,

---

\(^6\) Thymine (T) and Cytosine (C) belong to the group of pyrimidine.

\(^7\) Adenine (A) and Guanine (G) belong to the group of purine.
of a quantity $\Delta y(l)$ defined by $\Delta y(l) \equiv y(l_0+l) - y(l_0)$. Here the bars indicate an average over all positions $l_0$ in the gene.

\[
F^2(\bar{\Delta}) = [\bar{\Delta y(\bar{\Delta})}]^2 = [\Delta y(\bar{\Delta})]^2 - \Delta y(\bar{\Delta})^2,
\]

Fig. 6 The DNA walk representations of intron-rich human $\beta$-cardiac myosin heavy-chain gene sequence (a) its cDNA (b), and the intron-less bacteriophage $\lambda$ DNA sequence (c). Note the more complex fluctuations for the intron-containing gene in a compared with the intron-less sequences in b and c. Heavy bars denote coding regions of the gene. So that the graphical representation was not affected by the global differences in concentration between purines and pyrimidines, DNA walk representations were plotted so that the end point has the same vertical displacement as the starting point (for the statistical analysis, we use the original definitions (1)-(3), without any adjustment of vertical displacement). The minimum (min) and maximum (max) points on the landscape are denoted by arrows, and their application in the analysis is described in the text. For almost all intron-less genes and cDNA sequences studied, there appear regions with one strand bias, followed by regions of a different strand bias. The fluctuation on either side of the overall strand bias is random, a fact that is plausible by visual inspection of the DNA walk representations. (taken from [6])
3.7 Huffman coding [7]

Genome sequence analysis presents many difficult problems for scientists. The obstacles involved in the sequencing process, for example, include dealing with large amounts of data, lacking a complete knowledge of the genome length \( a \ priori \), and recognizing nucleotide symbol identity with complete accuracy. These impediments are typical of ones encountered in standard telecommunications problems.

By using a quaternary, real-valued DNA numerical sequence, the strings can be analyzes via the standard, lossless Huffman encoding technique. For the \( k \)th element, \( x[k] \), of the sequence \( X \), we denote \( x[k] = \gamma_1 \) for A, \( x[k] = \gamma_2 \) for T, \( x[k] = \gamma_3 \) for C, and \( x[k] = \gamma_4 \) for G. The Huffman encoding process is performed on \( X \). This numerical designation allows for the efficient computation of occurrence probabilities of nucleotide triplets within the sequence, correlations among other nucleotides, and probable locations of nucleotide combinations within the entire genome.

Working in the encoded domain will allow for the further reduction of analytical complexity if the sequences are very long. For a source symbol \( \gamma_i \), we have code word \( K(\gamma_i) \) occurring with probability \( \pi_i \), where \( K \) is the coding of the source. The \( n \) code words’ average length is given by

\[
L = \sum_{i=1}^{n} d_i \pi_i
\]

where \( d_i \) is the length of each individual code word. In general, \( L \) is larger than the symbol length of the original sequence, but the total number of code words will be less. The human \( \beta \)-globin intergenomic sequence (Accession HUMHBB), of length \( N = 73,308 \), which is studied in is addressed here. Accordingly, the Huffman encoding algorithm on this sequence reduces the number of symbols from \( N = 73,308 \) in the sequence domain to \( N = 20,841 \) in the encoded domain.

Although the codebook generated for the Huffman encoder is not unique for each sequence, knowing the symbol probabilities, and necessarily the codebook, \( a \ priori \) allows for a uniquely decodeable sequence. Determining the correlations of the symbols in the encoded domain has not proven to be extremely useful, mainly because the codebook is not unique for each sequence. However, this transformation allows us to visualize DNA sequences from a new perspective. Consequently, this technique is worthy of mention not only because it hints at the value of the information theoretic techniques to study DNA sequences, but compressing and then exploring symbolic strings from a digital communications perspective is applicable to DNA data.
4 Typical Topics

4.1 Gene Prediction: Identification of Protein Coding DNA Regions

4.1.1 Characteristics of protein coding DNA regions [8]

It is well-known that base sequences in the protein-coding regions of DNA molecules have a period-3 component because of the codon structure involved in the translation of base sequences into amino acids. For eucaryotes (cells with nucleus) this periodicity has mostly been observed within the exons and not within the introns. There are theories explaining the reason for such periodicity, but there are also exceptions to the phenomenon.

4.1.2 DNA Filtering Example: IIR Antinotch Filter [8]

To perform gene prediction based on the period-3 property, one defines indicator sequences for the four bases and computes the DFT’s of short segments of these, as described in section 3.1. The DFT of a length-$N$ block of $x_A(n)$ is defined as

$$X_A[k] = \sum_{n=0}^{N-1} x_A(n)e^{-j2\pi kn/N}, \quad 0 \leq k \leq N - 1$$

where we have assigned the number $n = 0$ to the beginning of the block. The DFT’s of other bases are defined similarly. The period-3 property of a DNA sequence implies that the DFT coefficients corresponding to $k / N = 3$ are large. Thus if we take $N$ to be a multiple of 3 and plot then we should see a peak at the sample value $k / N = 3$ as demonstrated in many papers. While this is generally true, the strength of the peak depends markedly on the gene. It is sometimes very pronounced, sometimes quite weak. Notice that a calculation of the DFT at the single point $k / N = 3$ is sufficient. The window can then be slided by one or more bases and $S[N/3]$ recalculated. Thus, we get a picture of how $S[N/3]$ evolves along the length of the DNA sequence. It is necessary that the window length $N$ be sufficiently large (typical window sizes are a few hundreds, e.g., 351, to a few thousands) so that the periodicity effect dominates the background $1/f$ spectrum (another characteristic of $S[N]$ which is not to be discussed in this tutorial). However a long window implies longer computation time, and also compromises the base-domain resolution in predicting the exon location. The use of IIR antinotch
filters for gene prediction was proposed by P. P. Vaidyanathan et al. A summary of the method is provided in section 3.1 of [8].

4.1.3 DNA spectrogram [1]

It is well known that the appearance of spectrograms provides significant information about signals, to the extent that trained observers can figure out the words uttered in voice signals by simple visual inspection of their spectrograms. Similarly, it appears that spectrograms are powerful visual tools for biomolecular sequence analysis. Here a proof-of-concept discussion defining a spectrogram as the display of the magnitude of the short-time Fourier transform (STFT), using the discrete Fourier transform (DFT) as a simple example of a frequency-domain analysis tool is presented.

Here the method of indicator sequence is utilized, and the spectrum defined in (4) is adopted, with a modification concerning reduction of dimensionality such that the sequences given in (6), namely

\[
x = \frac{\sqrt{2}}{3}(2u^l - u^c - u^g), \quad y = \frac{\sqrt{6}}{3}(u^c - u^g), \quad z = \frac{1}{3}(3u^d - u^l - u^c - u^g),
\]

are used.

The spectrograms of biomolecular sequences that simultaneously provide local frequency information for all four bases is defined by displaying the resulting three magnitudes by superposition of the corresponding three primary colors, red for \(x\), green \(y\), and blue for \(z\). Thus, color conveys real information, as opposed to pseudocolor spectrograms, in which color is used for contrast enhancement. For example, Fig. 7 shows a spectrogram using DFT’s of length 60 of a DNA stretch of 4,000 nucleotides from chromosome III of C. elegans (GenBank Accession number NC000967). The vertical axis corresponds to the frequencies \(k\) from 1 to 30, while the horizontal axis shows the relative nucleotide locations starting from nucleotide 858,001; only frequencies up to \(k = 30\) are shown due to conjugate symmetry as \(x\), \(y\), and \(z\) are real sequences. The DNA stretch contains three regions (C. elegans telomere-like hexamer repeats) at relative locations (953-1066), (1668-1727), and (1807-2028). These three regions are well depicted as bars of high-intensity values corresponding to the particular frequency \(k = 10\) (because period 6 corresponds to \(N/6 = 10\)). Other frequencies also appear to play a prominent role in the whole region of the 4,000 nucleotides. For comparison purposes, Fig. 8 shows the texture of a spectrogram coming from a sample of totally

---

random DNA, i.e., in which each type of nucleotide appears with probability 0.25 and independent of the other nucleotides.

Of course, there are numerous other ways in which spectrograms can be defined. We may use tapered windows, and adjust their width and shape. Furthermore, more balanced spectrograms can be defined using the wavelet transform rather than the DFT. The wavelet transform has been used to analyze some fractal scaling properties of DNA sequences.

Some remarks: The spectrograms and color maps introduced in this paper do not even search for obvious clues in the “character string domain,” like start codons, open reading frames, and splice site nucleotide pairs. They are meant to complement the existing tools that directly look into character string patterns, satisfying the need for direct visualization of certain local properties of biomolecular sequences.

Fig. 7  Color spectrogram of a DNA stretch. [1]

Fig. 8  Color spectrogram of “totally random” DNA. [1]
4.1.4 Identification of Protein Coding DNA Regions [1, 9, 10]

An example demonstrating the period-3 property of coding DNA sequences is shown in Fig. 9, where a coding region of length $N=1320$ inside the genome of the baker’s yeast (formally known as $S. \text{cerevisiae}$) demonstrates a peak at frequency $k = 440$. If we define the following normalized DFT coefficients at frequency $k = N/3$:

$$W = \frac{1}{N} X \left[ \frac{N}{3} \right]$$

$$A = \frac{1}{N} U_A \left[ \frac{N}{3} \right], \quad T = \frac{1}{N} U_T \left[ \frac{N}{3} \right],$$

$$C = \frac{1}{N} U_C \left[ \frac{N}{3} \right], \quad G = \frac{1}{N} U_G \left[ \frac{N}{3} \right]$$

then it follows from section 3.3 (spectral envelope approach), with $k = N/3$, that:

$W=aA+tT+cC+gG$. In other words, for each DNA segment of length $N$ (where $N$ is a multiple of three), and for each choice of the parameters $a$, $t$, $c$ and $g$, there corresponds a complex number $W=aA+tT+cC+gG$, which is a random variable.

For properly chosen values of $a$, $t$, $c$ and $g$, the magnitude of $W$ is a superior predictor, compared to $S[N/3]$, the coefficient in the total spectrum, of whether or not the DNA segment is part of a protein coding region; and that, in the former case, the phase $\Theta = \arg\{W\}$ is a powerful predictor of the reading frame that it belongs.

The chromosome XVI of $S. \text{cerevisiae}$ (GenBank accession number NC 001148) is considered here. All genes for which there were no introns and for which the evidence was labeled “experimental” are isolated. It is found that, for that particular chromosome, the average values of $A$, $T$, $C$, and $G$, scaled by $10^3$, were $8.0-56.3j$, $-84.1+37.4j$, $-46.2-23.2j$, and $122.3+42.1j$. By comparison, the magnitudes of $A$, $T$, $C$, and $G$, for nonprotein coding regions are much smaller, typically between one and two. The result of the proposed method using $|W|^2$ is shown in Fig. 10, and it is validated by Table 2. Detailed deduction of the method is omitted here, interested readers are referred to [1], [9], or [10].
Fig. 9  Plot of the spectrum of a coding DNA region, demonstrating peak at frequency $k = N/3$. [1]

Fig. 10  Plot of $|W|^2$ for the five exons shown in Table 2. [1]
### 4.2 Identification of Reading Frame [1, 9]

As mentioned in the preceding subsection, the phase of \( W \) is predictive of reading frame. The reason is that different reading frames exhibit different statistical characteristics. The angles \( \varphi_1, \varphi_2, \) and \( \varphi_3 \) are defined to be the expected values of the phase of the random variable \( W \) corresponding to the reading frames 1, 2, and 3, respectively. It is found that \( \text{mod}(\varphi_2 - \varphi_1) = \text{mod}(\varphi_3 - \varphi_2) = \text{mod}(\varphi_1 - \varphi_3) = -2\pi/3 \). To maximize predictive power, it is desirable to select the parameters \( a, t, c, \) and \( g \) minimizing some measure of the variability (such as the statistical variance) of \( \Theta = \arg\{W\} \). The data are normalized such that \( E\{\Theta\} = 0 \), and the \( \Theta \)'s in each STFT window can be color coded for visualization and reading frame identification. (As in the preceding subsection, detailed deduction of the method is omitted here, interested readers are referred to [1], [9], or [10].)

<table>
<thead>
<tr>
<th>Relative Position</th>
<th>Exon Length</th>
<th>Reading Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>929-1135</td>
<td>207</td>
<td>2</td>
</tr>
<tr>
<td>2528-2857</td>
<td>330</td>
<td>2</td>
</tr>
<tr>
<td>4114-4377</td>
<td>264</td>
<td>1</td>
</tr>
<tr>
<td>5465-5644</td>
<td>180</td>
<td>2</td>
</tr>
<tr>
<td>7255-7605</td>
<td>351</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2  Locations and Reading Frames of the Five Exons of the Gene F56F11.4. [1]

### 4.2.1 Color coding and color map approach [1, 9]

Because the number of primary colors (red, green, and blue) is the same as the number of possible forward coding reading frames, we can conveniently assign a color-coding scheme in which the value \( \Theta=0^\circ \) is assigned the color red, the value \( \Theta=120^\circ \) is assigned the color blue, and the value \( \Theta=-120^\circ \) is assigned the color green. In-between values are color-coded in a linear manner, according to Fig. 11, in which the three axes labeled \( R, G, \) and \( B \) correspond to the primary colors red, green, and blue.

The above color coding is used for reading frame identification, as shown in Table 3.

All STFT windows must be aligned at the same reading frame. Therefore, the sliding window should slide by precisely three locations for each DFT evaluation.
Furthermore, the window size should always be a multiple of three so that the frequency $k = N/3$ is well defined. Fig. 10 identifies the five exons based on the magnitude of the STFT using some specific parameter values of $a$, $t$, $c$, and $g$. Now different set of parameter values are used to enrich the information of Fig. 10 in the form of a color map shown in Fig. 12. For each nucleotide location in the color map, the color assigned obeys the rule of Fig. 11, and the intensity is modulated by the square-magnitude multiplied by 700 and clipped to the interval $(0, 1)$.

Note that the color of the third exon is closer to orange than to pure red, but the information is still sufficient to accurately identify its reading frame as 1.

![Fig. 11 Color coding of the Fourier transform phase. [1]](image)

<table>
<thead>
<tr>
<th></th>
<th>Reading frame 1</th>
<th>Reading frame 2</th>
<th>Reading frame 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Reading frame 1</td>
<td>Reading frame 2</td>
<td>Reading frame 3</td>
</tr>
<tr>
<td>Green</td>
<td></td>
<td>Reading frame 2</td>
<td>Reading frame 3</td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>Reading frame 2</td>
<td>Reading frame 3</td>
</tr>
</tbody>
</table>

Table 3 Color-Coded Reading Frame Identification. [1]
4.3 Prediction of Gene Function [4, 8]

In the previous sections, the capability of obtaining DNA and protein spectrum and the relation between the two (refer to section 3.0) has been demonstrated. When the spectrums of proteins with similar functions (say, hemoglobin of several species) are multiplied together and the magnitude of the spectrum is taken, a consensus spectrum is obtained. Through this and the relation of protein spectrum to DNA spectrum, it is possible to predict the function of genes identified in novel DNA sequences.

4.4 Long-range correlation [6]

Using DNA walk approach, the calculation of $F(l)$ can distinguish three possible types of behavior. (i) If the base pair sequence were random, then its root-mean-square value $C(l)$ would be zero on average (except $C(0)=1$), so $F(l)\propto l^{1/2}$ (as expected for a normal random walk). (ii) If there were a local correlation extending up to a characteristic range $R$ (such as in Markov chains), then $C(l)\propto \exp(-l/R)$, and for finite values of $l$ the $F(l)$ function would significantly deviate from $l^{1/2}$; nonetheless the asymptotic behavior $F(l)\propto l^{1/2}$ would be unchanged from the purely random case. (iii) If there is no characteristic length (i.e., if the correlation were “infinite-range”), then the scaling property of $C(l)$ would not be exponential, but would most likely to be a power-law function, and the fluctuations will also be described by a power-law

$$F(l)\propto l^\alpha$$

with $\alpha \neq \frac{1}{2}$.

Fig 6a shows a typical example of an intron-containing gene. The DNA walk has an obviously very jagged contour which corresponds to long-range correlations. The
calculation of $F(l)$ for this gene shows that the data are linear over three decades on this double logarithmic plot, which confirms that $F(l) \propto l^\alpha$. The least-squares fit yields the slope $\alpha = 0.67 \pm 0.01$.

4.5 Study on Gene Regulation

A magical interplay between proteins and DNA is responsible for many of the essential processes inside all living cells. Typically, each gene is being activated or expressed (starting the process that will eventually lead to protein synthesis) as a result of the combined presence, or absence, of certain particular regulatory proteins which bind to specific sites belonging to regulatory regions in DNA (usually in the vicinity of the gene) in a sequence-specific manner. DNA regulatory regions can be as short as ten nucleotide pairs in simple organisms, but can be thousands of nucleotide pairs in more advanced organisms; these nucleotide pairs store some complex digital logic involving chemical binding to complexes of multiple molecules, including several regulatory proteins. Again, chemical binding is dependent on the sequence-specific, 3-D structure of the macromolecules. Deciphering this digital logic in regulatory regions has proved to be a much more challenging task compared to the discovery of the genetic code governing coding DNA regions. We still know very little about these sophisticated regulatory mechanisms that govern the rates of activations of each of the genes.

Things become more complex, and more interesting, by the fact that each of the regulatory proteins are synthesized from other genes, which in turn were activated in relation to another set of regulatory proteins, and so on. A complex system can be defined by a network of many mutually interacting genes; the chemical product of each of these genes influences the activation of other genes in the network. One way of attempting to model this system is by using a set of nonlinear, differential equations involving concentrations of several proteins and other molecules that participate in related pathways. The output of such a system is a script involving the coordinated activation events of many genes; the precise timing of several such events during the lifecycle of the cell plays a crucial role. Even referring to primitive organisms, the term Bacterial Nanobrain has already been used to describe such networks which are indeed described as complex, generalized, artificial neural networks. Such gene regulatory networks are in the heart of genomic information processing, and their analysis is one of the most exciting future topics of research that will require a systems-based approach involving cross-disciplinary collaboration at various levels of abstraction, including a genomic level, a macromolecular binding level, and a higher network level.
5 Concluding Remarks

Signal processing-based computational and visual tools are meant to synergistically complement character-string-domain tools that have successfully been used for many years by computer scientists. In this tutorial, some of several possible ways that signal processing can be used to directly address biomolecular sequences are illustrated. The assignment of optimized, complex numerical values to nucleotides (as described in section 4.1 to 4.2) and amino acids provides a new computational framework, which may also result in new techniques for the solution of useful problems in bioinformatics, including sequence alignment, macromolecular structure analysis, and phylogeny.

An important advantage of DSP-based tools is their flexibility. Spectrograms can be defined in many ways. For example, depending on the particular features that must be emphasized, we may wish to define spectrograms using certain values of parameters. Once a visual pattern appears to exist, we have the opportunity to interactively modify the values of these parameters in ways that will enhance the appearance of these patterns, thus clarifying their significance. It is hoped that visual inspection of spectrograms will establish links between particular visual features (like areas with peculiar texture or color) and certain yet undiscovered motifs of biological sequences.

With the explosive growth of the amount of publicly available genomic data, a new field of computer science, bioinformatics, has emerged, focusing on the use of computers for efficiently deriving, storing, and analyzing these character strings to help solve problems in molecular biology. A plethora of computational techniques familiar to the signal processing community has already been used extensively and with significant success in bioinformatics, including such tools as hidden Markov models and neural networks. This is another area in which DSP-based approaches can be of help.

Gene regulation analysis is one of the most exciting research topics that can potentially be addressed using the theory of artificial neural networks. One of the tools providing valuable information about gene expression patterns is the DNA hybridization microarray. It is believed that there exists a unique opportunity for the DSP community, and the electrical engineering community in general, to play an important role in the emerging field of genomics.
6 Reference


Bioinformatics

The science of managing and analyzing biological data using advanced computing techniques. Especially important in analyzing genomic research data.

cDNA (complementary DNA)

DNA that is synthesized in the laboratory from a messenger RNA template.

Codon

A codon is a trinucleotide sequence of DNA or RNA that corresponds to a specific amino acid. The genetic code describes the relationship between the sequence of DNA bases (A, C, G, and T) in a gene and the corresponding protein sequence that it encodes. The cell reads the sequence of the gene in groups of three bases. There are 64 different codons: 61 specify amino acids while the remaining three are used as stop signals.

DNA (Deoxyribonucleic Acid)

DNA is the chemical name for the molecule that carries genetic instructions in all living things. The DNA molecule consists of two strands that wind around one another to form a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases-adenine (A), cytosine (C), guanine (G), and thymine (T). The two strands are held together by bonds between the bases; adenine bonds with thymine, and cytosine bonds with guanine. The sequence of the bases along the backbones serves as instructions for assembling protein and RNA molecules.

DNA Sequencing

DNA sequencing is a laboratory technique used to determine the exact sequence of bases (A, C, G, and T) in a DNA molecule. The DNA base sequence carries the information a cell needs to assemble protein and RNA molecules. DNA sequence information is important to scientists investigating the functions of genes. The technology of DNA sequencing was made faster and less expensive as a part of the Human Genome Project.

Exon

An exon is the portion of a gene that codes for amino acids. In the cells of plants and animals, most gene sequences are broken up by one or more DNA sequences called introns. The parts of the gene sequence that are expressed in the protein are called exons, because they are expressed, while the parts of the gene sequence that are not expressed in the protein
are called introns, because they come in between--or interfere with--the exons.

**Gene**

The gene is the **basic physical unit of inheritance**. Genes are passed from parents to offspring and contain the information needed to specify traits. Genes are arranged, one after another, on structures called chromosomes. A chromosome contains a single, long DNA molecule, only a portion of which corresponds to a single gene. Humans have approximately 23,000 genes arranged on their chromosomes.

**Gene regulation**

Gene regulation is the process of **turning genes on and off**. During early development, cells begin to take on specific functions. Gene regulation ensures that the appropriate genes are **expressed** at the proper times. Gene regulation can also help an organism respond to its environment. Gene regulation is accomplished by a variety of mechanisms including chemically modifying genes and using regulatory proteins to turn genes on or off.

**Genome**

The genome is the **entire set of genetic instructions** found in a cell. In humans, the genome consists of 23 pairs of chromosomes, found in the nucleus, as well as a small chromosome found in the cells’ mitochondria. These chromosomes, taken together, contain approximately 3.1 billion bases of DNA sequence.

**Genomics**

Genomics refers to the **study** of the entire genome of an organism whereas genetics refers to the study of a particular gene.

**Intron**

An intron is a portion of a gene that **does not code for amino acids**. In the cells of plants and animals, most gene sequences are broken up by one or more introns. The parts of the gene sequence that are expressed in the protein are called **exons**, because they are expressed, while the parts of the gene sequence that are not expressed in the protein are called introns, because they come in between the exons.

**Motif**

A **conserved element** of a protein sequence alignment that **usually correlates with a particular function**. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be **predictive** of any subsequent occurrence of such a structural / functional region in any other novel protein sequence.
Nucleotide

A nucleotide is the **basic building block of nucleic acids**. RNA and DNA are polymers made of long chains of nucleotides. A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base. The bases used in DNA are **adenine (A)**, **cytosine (C)**, **guanine (G)**, and **thymine (T)**. In RNA, the base **uracil (U)** takes the place of thymine.

Protein

Proteins are an important class of molecules found in all living cells. A protein is composed of one or more **long chains of amino acids**, the sequence of which corresponds to the **DNA sequence of the gene that encodes it**. Proteins play a variety of roles in the cell, including structural (cytoskeleton), mechanical (muscle), biochemical (enzymes), and cell signaling (hormones). Proteins are also an essential part of diet.

Reading frame

A sequence of codons beginning with an **initiation codon** and ending with a **termination codon**, typically of at least 150 bases (50 amino acids) coding for a polypeptide or protein chain.

Glossaries in bioinformatics can be found in:

- **Genome.gov | Talking Glossary of Genetic Terms**
  http://www.genome.gov/glossary/index.cfm
- **Glossary of Bioinformatics Terms**
- **Bioinformatics glossary**
  http://falcon.roswellpark.org/labweb/glossary.html
- **BioTech FYI Center – Glossary**
  http://biotech.fyicenter.com/glossary/
- **Appendix 1 A glossary of bioinformatics terms**
  http://www.bioinfbook.org/appendix.php