Last modified August 13, 2018

Elementary Sequence Analysis

Brian Golding, Dick Morton and Wilfried Haerty

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website http://helix.biology.mcmaster.ca/courses.html. Some of the programs that you will be using in this course and which will be run locally can be found at http://evol.mcmaster.ca/p3S03.html.

The "blue text" should designate links within this document while the "red text" designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo,  , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover . . .

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- methods to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- methods for detecting patterns and codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA Morton@McMaster.CA HaertyW@McMaster.CA

This document was entirely constructed with pdfLaTeX. Enormous kudos to all those that make this great software free to everyone.

Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

Contents

Chapter 10

Pattern Analysis

What is "random"? Intuitively, our idea of randomness is closely connected with homogeneity. Properties of a random sequence should somehow look the same at different scales. If they don't, we describe the sequence as "patchy". All genomes are complex and patchy. Some examples of DNA sequence heterogeneity are protein-coding regions, introns, CpG islands and dispersed tandem repeats such as the 171 human alpha satellite repeat.

What forces create heterogeneity in DNA sequences? Mutation is often though of as random. However, it is a complex process that does not occur uniformly across a genome. The process of replication, for example, may favor the expansion of repetitive regions by slippage. Transcriptionally active DNA may be subject to different mutational forces than non-transcribed regions. Regulatory elements may have different compositional requirements than coding regions. Natural selection is strong force creating DNA heterogeneity. Protein-coding regions experience complex selection intensities that vary among different codon positions and near splice junctions. Evolutionary history also affects sequence composition. Bacterial genomes are a mosaic of resident and horizontally transferred segments. Regions recently acquired from another organism with different base composition may appear as compositional heterogeneity.

10.1 Base Composition: first order patchiness

The fraction of bases that are G or C in a sequence varies dramatically among organisms. The range is greatest among bacterial taxa, which vary from about 30% to 75% (G+C). Genomes of mitochondria and chloroplasts tend to have higher (A+T) content then their host's nuclear genome as do introns compared to flanking exons. Causes for such variation are largely unknown although (G+C) content influences replication, transcription and translation through effects on secondary structure and the stability of double stranded molecules. Mutation and repair processes also affect DNA composition. It is tempting to speculate that higher (G+C) content is associated with thermostability since GC base pairs increase the melting temperature of DNA. However, there is no correlation between (G+C) and optimum growth temperature among prokaryotic genera (Galtier and Lobry, 1997 J. Mol. Evol. 44: 632).

10.1.1 Genome Patchiness

Differences in nucleotide composition are observed within genomes as well as between genomes. Karlin and Brendel (Science 259: 677-680, 1993) discussed the statistical analysis of DNA patchiness. Base content fluctuates at many different scales. One example is the large (>100 kb) regions in vertebrate genomes called "isochores" (Bernardi, Annu. Rev. Genetics 29: 445-476, 1995). Isochores are correlated with the staining properties of vertebrate chromosomes (Giemsa-positive and -negative bands). They have been revealed by physical analysis of DNA fragments as well as from DNA sequences (Ikemura *et al*., Genomics 8: 207-216, 1990). Genes tend to be concentrated in (G+C)-rich regions, but both coding and non-coding portions are subject to similar influences on composition. DNA sequence analysis of one isochore boundary indicated that it is sharp (Fukagawa *et al*., Genomics 25: 184-191; 1995). The origin of isochores is not clear. Bernardi favors an evolutionary explanation based on composition differences between warm and cold-blooded

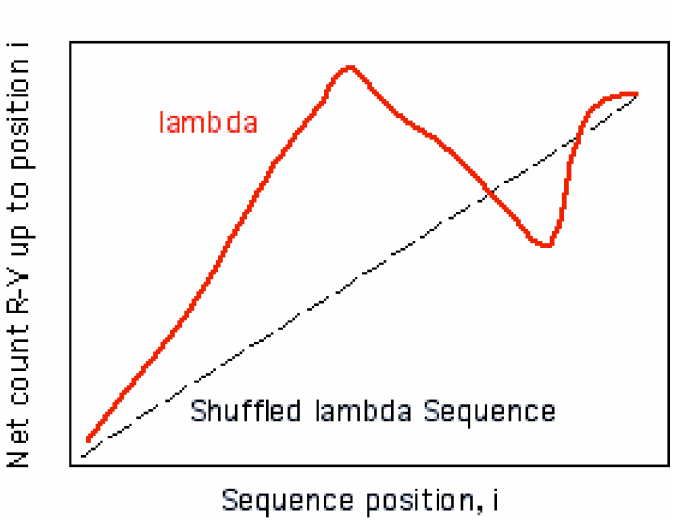


Figure 10.1: Sequence walk plot of the lambda genome after Karlin and Brendel, 1993.

animals. (G+C)-rich isochores are prominent in mammals and birds, although gene clustering and composition patchiness has also been observed in plants. Bernardi suggests that the (G+C)-rich isochores of mammals and birds originated about 200 million years ago from corresponding (G+C)-rich regions in their ancestors.

How can compositional patches be detected and what do they mean? These are questions that are actively pursued but not satisfactorily answered. Sequence walks are a simple method used to detect patchiness (Karlin and Brendel, 1993). As position is increased along a DNA sequence, the value of a variable is incremented +1 or 1 depending on a compositional parameter. Figure 10.1 is a sequence walk plot for the bacteriophage lambda genome where +1 is taken if the position is A or G (R=purine) and -1 if T or C (Y=pyrimidine) as described by Karlin & Brendel (1993). A randomly shuffled lambda sequence shows a steady increase in R-Y, while the actual lambda sequence has a patchy distribution of purines and pyrimidines.

Patchiness can also be visualized using a sliding window approach. Compositional parameters such as the (A+T) fraction are evaluated within a window that slides along the DNA sequence. Figure 10.2 is an (A+T) plot for the *E. coli* K12 genome. No unusual features are revealed in spite of the fact that the K12 chromosome contains several horizontally transferred regions.

10.2 Dinucleotide Composition: second order patchiness

Kornberg and his colleagues in the 1960s developed biochemical techniques for determining the dinucleotide content of DNA (Josse, J, Kaiser, AD and Kornberg, A, J. Biol. Chem. 261: 864-875, 1961). DNA is copied from a template using a 5′p labeled nucleoside triphosphate (pp*pY). The product is then cleaved with an enzyme that leaves a 3′p (Xp*). The radioactivity in Xp* is proportional to the amount of XpY in the DNA. Relative XpY values (nearest neighbor frequencies) are normalized by the amounts of X and Y to give a dinucleotide spectrum. These spectra were found to be characteristic of groups of organisms and were called the “general design” of DNA. They were used in cluster analysis to group organisms according to similarity in dinucleotide composition (Russell, GJ and Subak-Sharpe, JH, Nature 266: 533-536, 1977).

Karlin and his coworkers (Karlin, S, Campbell, AM and Mrázek, J, Annu. Rev. Genet. 32: 185-225, 1998) extended these biochemical methods to the computational analysis of DNA sequences. Following earlier work with “general design”, Karlin suggested that dinucleotide frequencies can be used as a “genome signature”. The normalized dinucleotide frequencies (called dinucleotide signatures) for a single DNA strand are given by equation 10.1 where f_{XY} is the frequency of XpY in the single strand and f_X is the frequency of X.

$$\rho_{XY} = f_{XY}/f_X f_Y \tag{10.1}$$

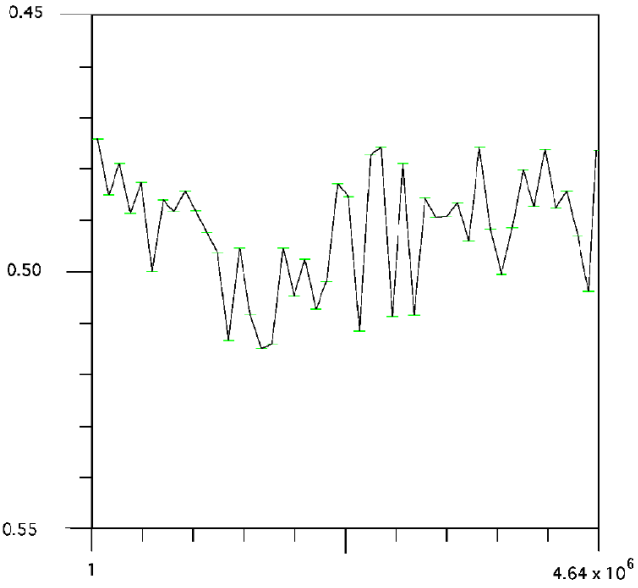


Figure 10.2: (A+T) fraction of the *E. coli* K12 chromosome, window = 100,000 nt

For normalized dinucleotide frequencies in dsDNA, the forward strand is concatenated with its complement in the above calculation. When the dinucleotide signature of XpY is >1.0, it is more frequent than expected from the nucleotide composition, while $\rho_{XY} < 1.0$ indicates under-representation. Karlin *et al.* (1998) suggest that $\rho_{XY} < 0.78$ or $\rho_{XY} > 1.23$ in 50 kbp or more of DNA are significant.

Genomic signatures may be useful for determining similarity within broad groups of organisms. They may also be able to detect horizontal transmission of DNA, provided the foreign DNA is from an organism with a different dinucleotide signature. For example, GpC dinucleotides are over-represented in the *E. coli* genome but not in some other bacteria such as *Pseudomonas*. There are many unexplained peculiarities about dinucleotide frequencies. For example, TpA is almost universally under-represented in DNA. Although this was observed in the biochemical studies of the 1960s, it has never been explained. The avoidance of CpG in vertebrate genomes is the one significant signature that has a theoretical basis. Vertebrates, but not invertebrates, methylate CpG (CpG $\rightarrow {}^{5m}$CpG). Deamination of ^{5m}C produces T so that ^{5m}CpG frequently mutates to TpG (mismatch repair is unable to correct TG pairs). Presumably, as CpG methylation evolved, the frequency of CpG dinucleotides decreased through mutation. With an important exception, CpG islands remain where methylation does not occur (Bird, AP, Nature 321: 209-213, 1986, see Figure 10.3). These unmethylated CpG islands are found in the 5′ regions of many genes, especially those that are constitutively expressed. Interesting, these CpG islands become hypermethylated in many tumors and gene expression is silenced (Esteller, M, Corn, PG, Baylin, SB and Herman, JG, Cancer Res. 61: 3225-3229, 2001). CpG methylation cannot be the complete story for the wide avoidance of this dinucleotide because CpG is also under-represented in mitochondrial genomes where it is not methylated.

10.3 Strand Asymmetry

10.3.1 Chargaff's Rules

Chargaff's rules express the fact that double stranded DNA obeys Watson-Crick base pairing. The two stands of dsDNA are sometimes labeled "Watson" and "Crick". Chargaff's first rules are $A_c = T_w$, $T_c = A_w$, $C_c = G_w$ and $G_c = C_w$, where the letters represent the molar fraction of a base on one strand. These rules result from formation of Watson- Crick base pairing between strands and are very precisely obeyed by dsDNA molecules.

Less well known are Chargaff's second rules. These apply only approximately and separately to each of the two strands

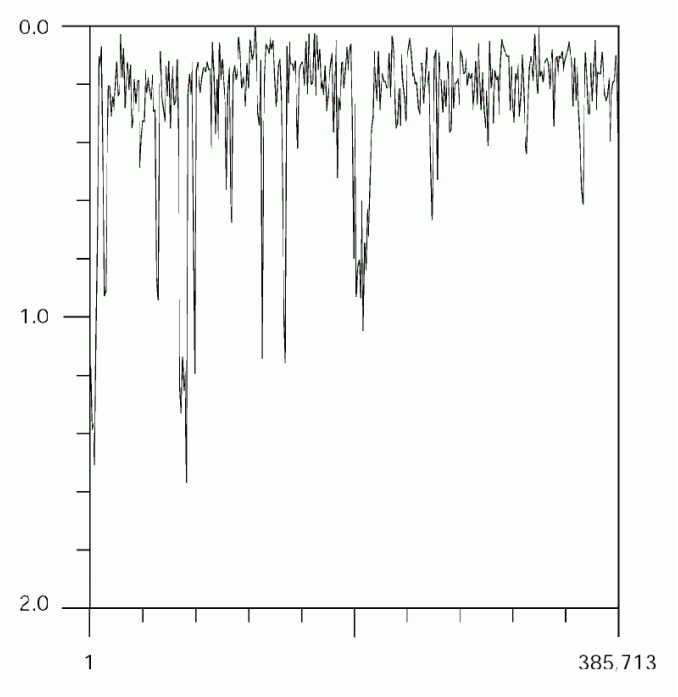


Figure 10.3: CpG islands in a 385 kbp segment of human DNA from chromosome 10 (Accession: AL031601). Dinucleotide signature (ρ_{GC} for CpG), window = 1,000 nt

of dsDNA. They are: $A_c \sim T_c$, $T_w \sim A_w$, $C_c \sim G_c$ and $G_w \sim C_w$. Chargaff's second rules express the fact that complementary strands are approximately symmetric in nucleotide content. If they are true, then $A_c = A_w$, $T_c = T_w$, $C_c = C_w$ and $G_c = G_w$. Departures from strand symmetry (Chargaff asymmetry) are expressed by differences: (A-T)/(A+T) and (G-C)/(G+C) on a single strand.

$$\begin{aligned} \phi_{AT} &= (f_A - f_T)/(f_A + f_T) \\ \phi_{GC} &= (f_G - f_C)/(f_G + f_C) \end{aligned} \tag{10.2}$$

Strand symmetry originates from identical substitution processes affecting each strand, for example, when changing $A_c \rightarrow T_c$ has the same probability as $A_w \rightarrow T_w$. Under these circumstances, the number of AT base pairs will approximately equal the number of TA base pairs (and likewise for GC and CG). However, some mutation processes are known to be strand asymmetric (Francino and Ochman, Trends Genet. 13: 240-245, 1997). Furthermore, nucleotide substitution is subjected to selection that may depend on information contained in only one strand.

10.3.2 Replication Asymmetry

The leading- and lagging-strands are replicated by different mechanisms. The leading-strand is copied by a continuous process, while the lagging strand is synthesized discontinuously using multiple, short RNA primers. Additional enzymes are needed to synthesize primers and then later remove them and fill in gaps. Leading- and lagging-strand replication may involve different polymerases with disparate error rates. As well, the structure of the replication fork exposes the leading- and lagging-strands to different environments. The lagging-strand is more open as a longer, single-stranded structure, which could lead to increased DNA damage and repair.

Mutagenesis experiments in *E. coli* have shown that deletions and replication errors are more frequent on the lagging strand. Differences depend on the agent inducing replication errors. Excess dTTP causes more errors on the lagging strand, while

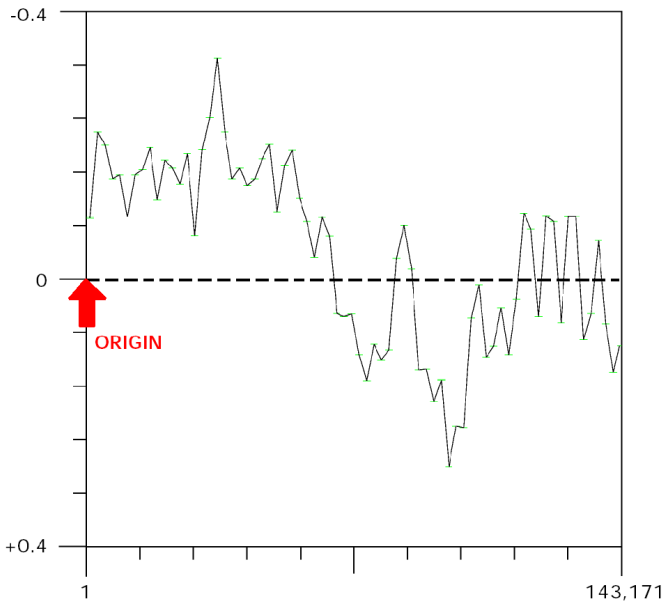


Figure 10.4: Strand asymmetry for the *Euglena gracilis* chloroplast chromosome (Accession: X70810) after Morton (1999). The chromosome is circular and strand asymmetry changes sign quickly at the replication origin and at a point about 180^0 from the origin. There are also peaks associated with the open reading frames and the three rRNA operons. $\phi_{AT} = (f_A - f_T)/(f_A + f_T)$ for window = 1,000 nt.

excess dCTP makes little difference. In general, it seems that $Y \geq R$ (pyrimidine $\geq$ purine) changes are more frequent on the lagging-strand, causing an accumulation of purines.

Replication bias may cause a switch in Chargaff asymmetry across a replication origin because at this point the leading- and lagging-strands change identity. An example is the *Euglena gracilis* chloroplast genome as reported by Brian Morton (Proc. Natl. Acad. Sci. USA. 96: 5123-5128, 1999), see Figure 10.4.

Lobry (Mol. Biol. Evol. 13: 660-665, 1996) analyzed the chromosomes of several bacteria for replication bias. The expected switch in strand asymmetry occurred across the replication origins. Changes in (G-C)/(G+C) were much more dramatic than changes in (A-T)/(A+T). The replication effect was partly obscured by protein-coding sequences, which introduce their own bias (see also the *Euglena* chromosome in Figure 10.4). Wherever one strand had a higher density of coding sequences, that strand was found to increase G>C and T>A. Contrary to the expectation from mutagenesis, the lagging-strand accumulated more A and C (instead of A and G).

No evidence has been found for replication bias in eukaryotes. Chargaff asymmetries switch rapidly over short regions of the chromosome although they are generally higher around protein-coding exons. Apparently, the effect of mutational bias and/or codon selection obscures the asymmetry (if any!) caused by a replication origin.

10.3.3 Transcriptional Asymmetry

Transcription can also introduce Chargaff asymmetry since the two strands may be subject to different mutational effects. During transcription, the non-template strand is in an open single-stranded conformation that is more sensitive to certain mutations such as C $\geq$ T (U) deamination. The template strand, on the other hand may be subject to transcription-dependent repair. DNA damage (for example a pyrimidine dimer) can stall the RNA polymerase and promote the action of nucleotide excision repair. This repair may be error-prone, inducing mutations on the template strand. Or unrepaired damage on the non-template strand may lead to substitution.

10.3.4 Codon Selection

Selection for specific amino acids in protein-coding DNA produces strand asymmetry. For example, suppose selection favors glycine in a protein. Thus, GGN codons tend to occur on one strand and complementary NCC nucleotides on the other. The content of G increases relative to C in the sense strand. Thus, protein amino acid composition can impose strand asymmetry. Certain kinds of codon bias in the synonymous position also produce strand asymmetry. One site to find a general description of DNA walks can be found at http://www2.unil.ch/comparativegenometrics/DNA_walk.html.

10.4 Simple Sequence Repeats

Runs of a repeated amino acid are common in the proteins of all organisms. The first triplet repeats, called "Opa", were discovered in the Drosophila Notch gene (Wharton, KA, Yedvobnick, B, Finnerty, VG, Artavanis-Tsakonas, S. Cell 40: 55-62, 1985). These are CAG (or CAA) repeats that code for glutamine (GLN = Q) when translated. Glutamine domains often form protein-protein inter- or intra-molecular contacts. They are an example of the general class of triplet repeats. CAG is the best-known, but CTG, GCC CGG as well as others are also common. Triplet repeats have an been associated with a number of genetic syndromes (Paulson, HL and Fischbeck, KH. Annu. Rev. Neurosci. 19: 79-107, 1996). They are not always found in protein-coding domains, but are also observed in non-coding sequences. They are a subset of minisatellite repeats that have been used for studies of DNA polymorphism, evolution and fingerprinting.

The reiteration of a single amino acid is only one way in which the complexity of protein- coding DNA is reduced. Brian Golding has found that "simple sequence" motifs are a common feature of proteins (Golding, GB. Protein Sci. 8: 1358-1361, 1999). Regions that contain repeated amino acids of varying complexity represent protein sequence simplification. An example is splicing factors that contain repeated "SR" (serine-arginine) domains. These domains are involved in protein-protein contacts that take place during dimerization.

Protein simplification can be detected by using information theory, which will be described in more detail in section 10.5.2. The Shannon-Weaver index is used as a measure of complexity. Figure 10.5 shows how information content reveals regions of low complexity in a yeast nuclear localization protein. The most dramatic is from about 10% to 30% of the protein sequence where reiterated serines frequently occur. Another region of high glycine content occurs at about 90% of the sequence.

10.5 Sequence Complexity

There are a number of ways that complexity in DNA or protein sequences might be represented. The best is based on information theory. Information theory describes the information content of a sequence of symbols. There is little information in repetitive symbols because the number of possible messages that can be made from them is small. On the other hand, sequences that appear random or complex can potentially contain a great deal of information.

10.5.1 Information Theory

Shannon and Weaver developed their theory of information in order to understand the transmission of electronic signals. Gatlin (Information Theory and the Living System. Columbia University Press, New York,1972) describes its extension to Biology. Information theory is an obvious tool to look for pattern and complexity in DNA and protein sequences (Schneider, 1995, Information theory primer. http://www-lmmb.ncifcrf.gov/ toms/paper/primer/primer.pdf). However, results from this area have so far been somewhat disappointing. Shannon and Weaver developed a measure for the information content of messages made from a sequence of elements, L elements in length, each element chosen from a set of (S_i) symbols with probability of occurrence p_i.

$$H = -L \sum p_i log_2(p_i) \qquad [log_2(p_i) = 1.4427 log_e(p_i)] \tag{10.3}$$

The units of H are called "bits". Since logarithms are additive, L in equation 10.3 can be removed (H/L) to give the

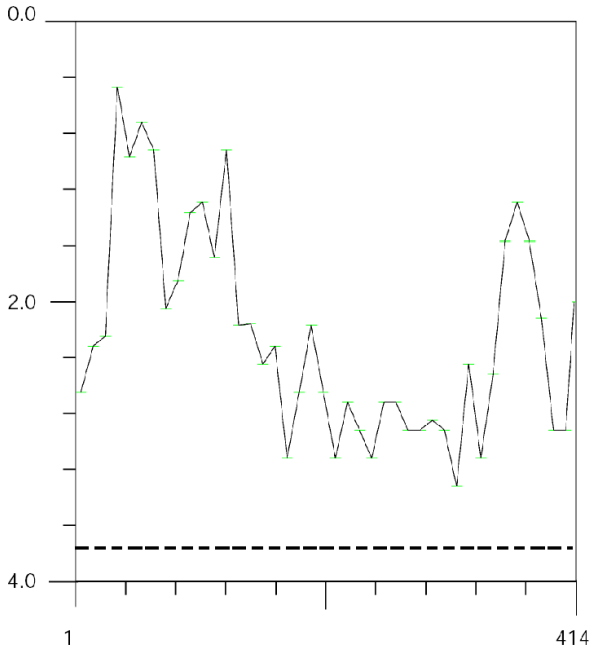


Figure 10.5: Amino acid complexity (Shannon-Weaver information content) in the *Saccharomyces cerevisiae* nuclear localization sequence binding protein; Nsr1p (Accession: NP_011675). The dashed line shows the Shannon-Weaver index for the entire protein sequence, the solid lines connect windows of 10 amino acids.

average value in bits per nucleotide (or amino acid) site. For a DNA sequence of length L containing four bases, the maximum entropy occurs when each of the four bases has equal frequency. In this case,

$$H_{max} = -L \quad 4 \quad \frac{1}{4} \quad log_2(\frac{1}{4})$$

So $H_{max} = 2L$ bits or 2 bits per nucleotide site. Each nucleotide site can be represented by a two bit number (11, 10, 01, 00). This is the maximum complexity of a DNA message. Less complexity is contained in sequences that depart from equal frequency. At the other extreme is a sequence comprised of a single base ($p_i = 1$, $H/L = 0$). The Shannon-Weaver index can be regarded as a measure of the complexity of a sequence. $H/L = 0$ represents a sequence of minimum complexity, $H/L = 2$ bits has maximum possible complexity.

One way to think about the Shannon-Weaver index is in terms of uncertainty. Suppose the four bases are equally likely. The uncertainty of a single base is 2 bits before it is read by a functional device (enzyme). After the base is decoded, its uncertainty is zero. The information content of the message is the decrease in uncertainty as a result of decoding.

Information theory has been applied to the analysis of DNA and protein sequences in three ways.

1. Analyzing sequence complexity from the Shannon-Weaver indices of smaller DNA fragments (windows) contained in a long sequence as was done in Figure 10.5.

2. Comparing homologous sites in a set of aligned sequences by means of their information content. That is, determining the complexity of homologous sites.

3. Examining the pattern of information content of a sequence divided into successively longer words (symbols) consisting of a single base pairs, triplets and so forth. This is a method to look at clustering of nucleotides and will not be considered.

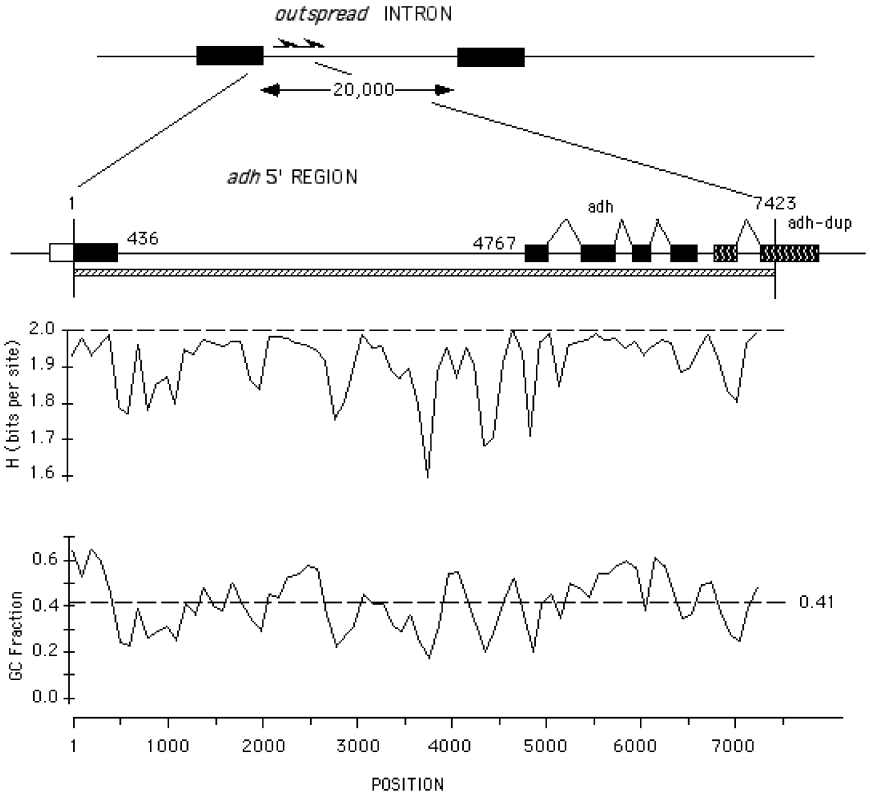


Figure 10.6: Nucleotide complexity (Shannon-Weaver information content H/L) of the *D. melanogaster* ADH gene region (Accession: Z00030), windows of 100 nucleotides overlapped by 50 nucleotides.

10.5.2 Sequence Window Complexity

An analysis of the *D. melanogaster* alcohol dehydrogenase (ADH) gene illustrates the application of information theory to DNA sequence data (Figure 10.6).

The ADH gene lies within a 20 kb intron of a larger gene, outspread. Generally, maximum complexity is found in exons of either ADH or outspread. In fact, the existence of the left-most exon in outspread was first deduced from an open reading frame 5′ to the ADH gene before the outspread gene had been mapped. Figure 10.6 also shows a correlation between complexity and base composition. In principle, increasing the relative frequency of any of the nucleotides should have the same effect, to decrease complexity. However, in this region of the *Drosophila* genome, only increased (A+T) decreases complexity, while increased (G+C) has the opposite effect. High GC is associated with protein-coding exons while high AT is associated with non-coding DNA such as introns. Although natural selection produces more constrained messages, proteins do not usually use highly patterned or repetitive codon choices except where simple amino acid repeats are found (see Figure 10.5). The Shannon-Weaver index reaches nearly the maximum value of 2 bits per site for the protein-coding exons of these two *Drosophila* genes. Regions of repetitive DNA, on the other hand, have low complexity. In the ADH region of the *Drosophila* genome, these are associated with AT-rich sequences. It is also interesting that intron DNA between ADH and outspread exons appears to be organized into sub-regions with different complexities. It remains to be seen if intronic regions of high complexity and GC content are functional and constrained by natural selection, as are protein-coding exons, or simply a different kind of neutral DNA.

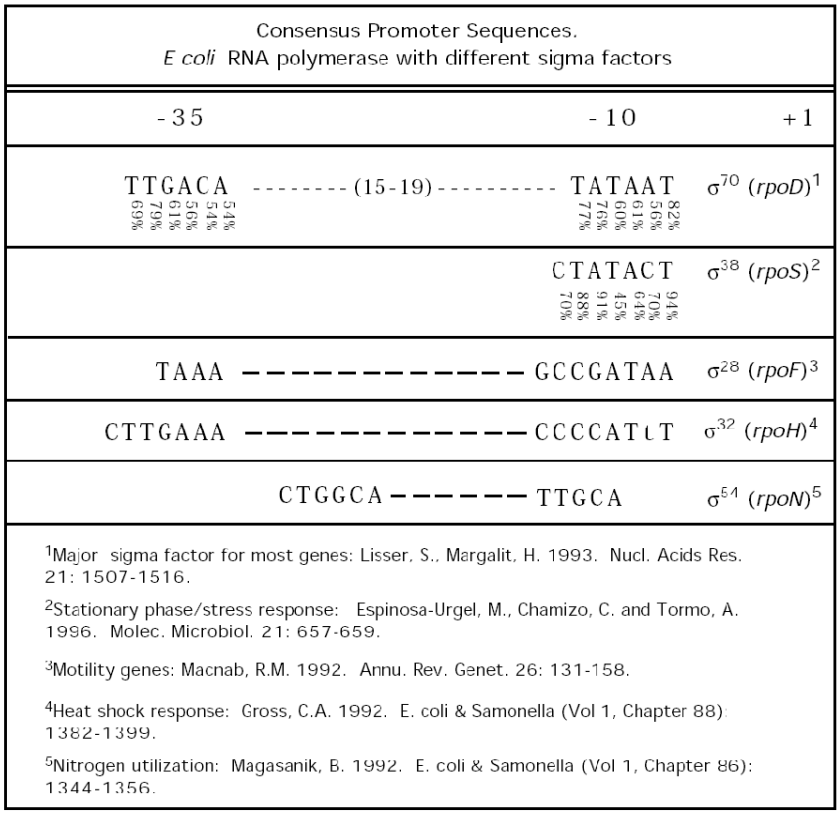


Figure 10.7: Consensus sequence analyses of *E. coli* promoters. +1 is the transcriptional start position.

10.6 Finding Pattern in DNA Sequences

A different and perhaps more important problem than compositional heterogeneity is the location of regulatory elements. Functionally important sequences are conserved across homologous DNA segments from different species. Conservation of DNA information is not restricted to evolution by descent. Convergence may produce similar patterns within a single genome. For example, DNA sequences recognized by the same or a similar DNA-binding protein will be conserved in order that the protein functions properly. I will describe a conserved sequence motif, the *E. coli* sigma70 promoter as an example of finding a conserved pattern within a genome.

10.6.1 Consensus Sequences

Promoter sequences, in conjunction with other DNA elements and proteins, activate RNA polymerase binding and transcription. *E. coli* promoter elements are recognized by an RNA polymerase holoenzyme which contains a bound sigma factor (core enzyme plus sigma factor = holoenzyme). The sigma factor is though to provide most of the sequence recognition capability of the holoenzyme. *E. coli* has a number of different sigma factors, each associated with a specific promoter consensus sequence (Figure 10.7).

The consensus sequnce is defined by majority rule. Analysis of the sigma-70 promoter by Lisser and Margalit (Nucleic Acids Res. 21: 1507-1516, 1993) revealed the consensus sequence shown in Figure 10.7. A pattern search for the sigma-70 promoter based on the consensus sequence would look for TTGACA $(N)_{n=15-19}$ TATAAT.

A major drawback to using the consensus sequence in pattern matching is that rarely will an actual promoter perfectly match the consensus sequence. No known sigma-70 promoter matches the consensus sequence at all 12 nucleotides (although this pattern does occur in the *E. coli* genome). Thus, searching for the consensus sigma-70 promoter sequence in front of

genes is an exercise in futility. The search must be for "something like" the consensus sequence. But how alike?

Variation found among the promoters of individual *E. coli* genes is indicated under the majority base in Figure 10.7. Some sites in the promoter sequence are more conserved than others. The cause of variation, however, is unknown. It could be due to mutational drift under the influence of selection. There may also be gene-specific effects. For example, genes requiring lower expression may use "weaker" promoters.

It is possible to take variation into account in a pattern search by defining alternative nucleotides. For example, if the most frequent alternative to the first T in TTGACA is A, the pattern search could be for (T/A)TGACA. Problems with simple pattern searches are obvious. The number of possible patterns grows exponentially with alternatives, but all of them are not equally useful as matches. A pattern with 10 mismatches from the consensus is probably not a promoter, but one with two mismatches might be. To account for this, pattern-matching programs will allow up to a specified number of mismatches. Another problem is that there may be no clear alternatives to the consensus nucleotide. This is the case with the *E. coli* sigma-70 promoter where minority nucleotides are more-or-less evenly distributed (Table 10.1).

Base	T	T	G	A	C	A	T	A	T	A	A	T
A	0.10	0.06	0.09	**0.56**	0.21	**0.54**	0.05	**0.76**	0.15	**0.61**	**0.56**	0.06
C	0.10	0.07	0.12	0.17	**0.54**	0.13	0.10	0.06	0.11	0.13	0.20	0.07
G	0.10	0.08	**0.61**	0.11	0.09	0.16	0.08	0.06	0.14	0.14	0.08	0.05
T	**0.69**	**0.79**	0.18	0.16	0.16	0.17	**0.77**	0.12	**0.60**	0.12	0.15	**0.82**

Table 10.1: Fractional occurrence of nucleotides at each position for 298 *E. coli* sigma-70 promoters (Lisser and Margalit, 1993)

10.6.2 Matrix Analysis of Sequence Motifs

Hertz and Stormo discuss the analysis and prediction of *E. coli* promoters (Methods in Enzymol. 273: 30-42, 1996). The basic method of analyzing sequence motifs and their conservation is to compute a score using a scoring matrix. The simplest scoring matrix assigns a score of one for each match with the consensus sequence (Figure 10.8). A perfect match to the consensus nucleotide produces the maximum score. Partial matches produce intermediate scores.

An assumption of using scoring matrices to evaluate potential sequence patterns is that each site must act independently. No covariance is allowed between nucleotide changes at one position with those at another position.

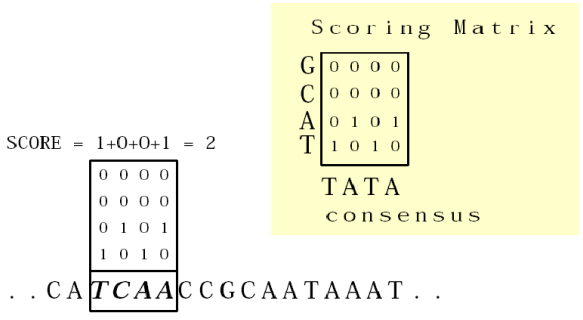


Figure 10.8: Matrix analysis of a sequence motif using a scoring matrix based on the consensus sequence (TATA). The score of 2 for TCAA indicates it matches the consensus sequence at two sites.

Scoring matrices can be developed that use more information about the pattern than contained in the consensus sequence. One approach is to find the matrix that gives the best correlation between the scores it produces and the measured activities of actual promoter sequences. If the activities of an example group of promoter sequences are equal, the maximum likelihood matrix elements will be the logarithms of observed frequencies for each nucleotide at a position divided by the probability that the nucleotide occurs by chance (equation 10.4). The latter can be estimated from the genome nucleotide frequency (e.g., p_i 0.25 for each base in *E. coli*).

$$W_{in} = \log_{10}(F_{in}/p_n) \tag{10.4}$$

W_{in} is the scoring matrix element at the i^{th} position in the pattern for the n^{th} type of nucleotide (G, C, A, or T); F_{in} is the frequency of the n^{th} nucleotide at the i^{th} position among the group of patterns used to derive the consensus sequence;

p_n is the probability that the nth nucleotide occurs by chance. For example, among the group of promoters used to derive the sigma-70 consensus sequence in Figure 10.7, the T at -10 (TATAAT) occurs 82% of the time (Table 10.1). The scoring element for a T at this position is

$$W_{TT} = \log_{10}(0.82/0.25) = 0.516 \tag{10.5}$$

(assuming that T occurs with a frequency of 1/4 in the *E. coli* genome).

Scores for DNA patterns can also be obtained using neural network methods. Examples of such techniques are discussed in Hénaut and Danchin (Analysis and predictions from Escherichia coli sequences, or *E. coli* in silico In: *E. coli* and *Salmonella* Vol. II, Chapter 114: 2047-2066, 1992). A computer program is "trained" on examples of good and bad promoters. Matrix elements are flexible and optimized to discriminate between the training set. Such methods do not usually give appreciably better results than the maximum likelihood approach. However, they can be more easily adapted to include additional information about what makes a good promoter. Many promoters require several proteins to initiate transcription. These recognize other DNA sequence motifs, usually located near the sigma factor binding site. DNA curvature is often a factor. Upstream sequences that bend DNA increase the activity of some promoters (Travers, Cell 60: 177-180, 1990). DNA bending depends mainly on runs of A or T since the dinucleotide AA/TT has the largest tilt angle (Trifonov, CRC Revs. Biochem. 19: 89-106, 1985). DNA curvature can be calculated by accumulating AA and TT pairs. DNA curvature is more easily incorporated into the analysis of promoter scores by using training methods.

10.6.3 Sequence Conservation and Sequence Logos

DNA regulatory elements such as promoter sequences are examples of a constraint placed on the evolution of DNA sequences by natural selection. Variability across genomes or among genes is reduced because a conserved protein molecule must recognize different forms of the element. Variability results from a balance between mutation and selection.

Information theory can be used to analyze the effectiveness of selection. Although the approach can be applied to any conserved DNA or protein sequence, its theoretical basis is clearest for DNA binding sites that are recognized by a protein molecule. In this case, the protein can be thought of as decoding information contained in its binding site. This information can be evaluated by comparing variation among different binding sites. Unlike the consensus sequence in which every position in the binding site is equivalent, information analysis evaluates the relative importance of individual sites. A good description of this method is the paper by Shaner, Blair and Schneider (1994, Sequence logos: a powerful, yet simple, tool. http://www-lecb.ncifcrf.gov/ toms/paper/hawaii/). A "sequence logo", is obtained from a set of aligned DNA (or protein) examples. The information content (R_i) of each site (i) is calculated from equation 10.6.

$$R_i = H_{max} - H_i - e(N) \tag{10.6}$$

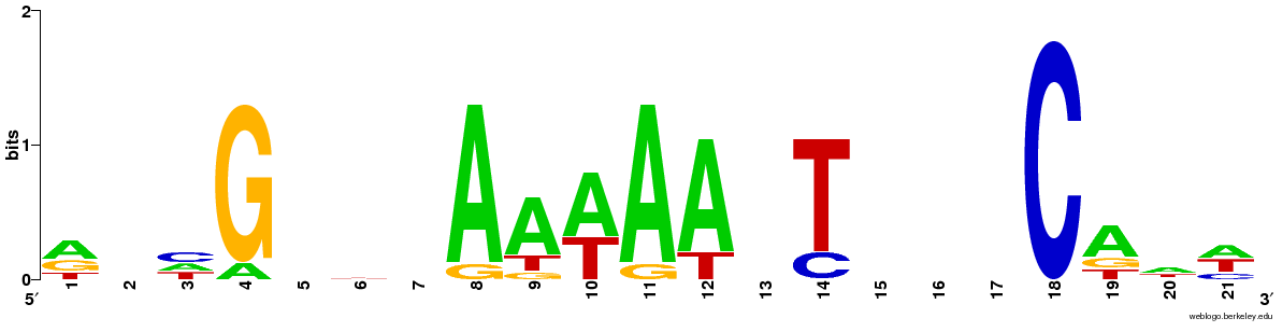
R_i is the information content of the site. H_{max} is the maximum uncertainty, 2 bits if the four bases are equally probable before the site is decoded (see section 10.5.1). After decoding (e.g., by the RNA polymerase for promoter sequences), the uncertainty (H_i) is given by equation 10.3 with p_i are nucleotide frequencies calculated from each position of the aligned, example sequences. e(N) is a correction factor to account for the fact that only a finite number of example sequences (N) are used to estimate the information content of the binding site (see Schneider et al, 1986). Figure 10.9 illustrates the method by analysis of the *E. coli* FIS binding site using data from Hengen et al (Nucleic Acids Res. 25: 4994- 5002, 1997).

F_{is} binds to and bends DNA at specific sites. It regulates the transcription of a subset of genes in conjunction with RNA polymerase, and is also involved in the process of recombination. F_{is} sites have been identified in a number of genes as well as the f_{is} gene itself. Some genes (e.g., f_{is}) have a cluster of sites in their promoter region. Figure 10.9 displays the information content of the F_{IS} binding site as a "sequence logo", where each consensus nucleotide is given a size proportional to its information. Hengen *et al.* analyze 60 example sequences (30 sites in both directions since the F_{is} site is known to be symmetrical) from which I selected 10 for illustration in Figure 10.9. Sequence logos can be constructed at the internet site: http://weblogo.berkeley.edu/.

An advantage of sequence logos is that sequence conservation can be quantitatively interpreted as the information that the decoder (e.g., F_{is} protein) obtains from potential sites in order to recognize a valid site. For example, the information content of the two GC base pairs in the F_{IS} binding site is approximately 2 bits, close to the maximum information available. The F_{IS} protein contacts the major groove of dsDNA at these positions and can obtain information about base pair identity (e.g., CG vs GC). On the other hand, in the central region of the F_{IS} binding site, the protein contacts the

Figure 10.9: Analysis of ten F_{IS} binding sites. The consensus is shown at top and the 'logo' at the bottom.

Consensus	A	A	C	G	C	T	C	A	A	A	A	A	T	T	G	A	C	C	A	A	A
fis	T	T	T	G	C	C	G	A	T	T	A	T	T	T	A	C	G	C	A	A	A
oriC	A	C	A	A	C	T	C	A	A	A	A	A	C	T	G	A	A	C	A	A	C
rrnB	A	A	C	G	G	G	C	A	A	T	A	A	T	T	G	T	T	C	A	G	C
tufB	G	A	T	G	T	T	G	A	A	A	A	A	G	T	G	T	G	C	T	A	A
tyrT	G	G	C	G	A	T	T	A	A	A	G	A	A	T	A	A	T	C	G	T	T
nrd	A	C	C	G	A	A	T	A	G	A	A	A	A	C	A	A	C	C	A	T	T
tgt	T	G	A	G	C	T	A	A	A	A	A	A	T	T	C	A	T	C	G	A	T
aldB	G	C	T	G	C	G	C	G	A	T	A	A	A	T	C	G	C	C	A	C	A
proP	A	A	A	G	G	T	C	A	T	T	A	A	C	T	G	C	C	C	A	A	T
hin	A	G	C	G	A	C	T	A	A	A	A	T	T	C	T	T	C	C	T	T	A



minor groove and can only distinguish GC from AT pairs, but not their orientation. The information available in this region is approximately 1 bit.

The information content of a binding site can be calculated by summing the information at each position. It is approximately 9 bits for the F_{is} consensus sequence (Hengen *et al*., 1997). This contrasts with a maximum of $21 \times 2 = 42$ bits of information available in a 21 bp binding site. The F_{IS} protein uses only a faction of the this information in order to recognize a site. Nine bits of information is sufficient to allow approximately 16,000 sites to be distinguished in the *E. coli* genome [$9 = -\log_2(x/G)$, where G is the genome nucleotide content, about 8×10^6 nucleotides because each nucleotide begins a potential site (see Schneider *et al*., 1986). The number of nucleotides in the *E. coli* genome was doubled because the F_{is} site is symmetric. Solution gives $x = 16,000$]. More stringent binding site recognition requires that more information to be used by the protein.

The total information of a potential binding site can be calculated using a scoring matrix derived from equation 10.7.

$$W_{bj} = H_{max} - \log_2(F_{bj}) - e(N) \tag{10.7}$$

W_{bj} is the matrix element for nucleotide of type b at position j in the pattern. F_{bj} is the frequency of this nucleotide in the example set at the same position, and e(N) is a correction for the finite size (N) of the example set. The information content of a test pattern is obtained by using its sequence in equation 10.7. Hengen *et al*. (1997) used this approach to scan the *E. coli* genome for F_{IS} binding sites. A sliding window of 21 nucleotides was moved along the genome sequence and the information content of potential sites evaluated. Segments with information above 2 bits were considered potential F_{IS} sites.