Mathematical Methods for DNA Sequences

Editor
Michael S. Waterman, Ph.D.
Professor of Mathematics and Molecular Biology
University of Southern California
Los Angeles, California

CRC Press, Inc.
Boca Raton, Florida

Chapter 5

SOME STATISTICAL ASPECTS OF THE PRIMARY STRUCTURE OF NUCLEOTIDE SEQUENCES

Simon Tavaré and Barton W. Giddings

TABLE OF CONTENTS

I. INTRODUCTION

The formidable volume of DNA sequence data generated in the last decade provides a rich source of data to biochemist, geneticist, and statistician alike. This paper studies some statistical aspects of the primary structure of nucleotide sequences.

We have divided our presentation into three main sections. The first gives a brief introduction to molecular biology of the nucleic acids, followed by some examples from the literature that give a feeling for the types of problems that are addressed. The second section describes some Markov chain methods for assessing the dependence structure that exists in a sequence of nucleotides. Particular emphasis is placed on methods for estimating the order of the Markov dependence. These methods are illustrated with an analysis of the bacteriophage λ genome.

The ordering of the bases in a nucleotide sequence is influenced by both random factors (such as mutation) and deterministic pressures (such as selection). The third part of our paper describes some methods for searching for repetitive or periodic patterns in a sequence. We base our analysis on the discrete Walsh transform and compare it to the more familiar Fourier methods.

Our aim has been to focus on some useful analysis techniques, without going into detail on all the variations on a given theme. The references will provide the interested reader with additional information, both biochemical and statistical, about this fascinating field.

II. STATISTICAL ASPECTS OF DNA SEQUENCES

A. Nucleic Acids — A Brief Review of Molecular Biology

The nucleic acids are of two types. The first, deoxyribonucleic acid (DNA), is composed of deoxyribonucleotides connected by phosphodiester linkages. There are four types of nucleotides in DNA: two purines, adenine (A) and guanine (G), and two pyrimidines, cytosine (C) and thymine (T). DNA may be single-stranded (as in certain single-stranded DNA viruses such as M13 or φX174 or when denatured by heat or alkali), but is usually found as double-stranded molecules. The strands of the double helix are held together by hydrogen bonding between bases on the two strands. A pairs with T form two hydrogen bonds per base pair, while G and C pair to form three bonds.

The second type of nucleic acid is ribonucleic acid (RNA). Like DNA, RNA is composed of nucleotides joined by phosphodiester bonds. The nucleotides of RNA, however, are ribonucleotides (ribose, the sugar component of each RNA nucleotide, has a hydroxyl group at the 2′ position, whereas deoxyribose, the sugar component of DNA, does not). Furthermore, uracil (U) replaces T in RNA molecules. RNA molecules are usually single-stranded, although there may be a great deal of intrastrand base pairing.

RNA occurs as one of three types of molecules. In a process known as "transcription", RNA polymerase makes a complementary copy of the genetic information (i.e., the base sequence) of one strand of DNA (called the "sense" strand) which is called messenger RNA (mRNA). The mRNA carries sequence information and serves as the template for the synthesis of proteins by a process known as "translation". Messenger RNA constitutes only about 5% of the total cellular RNA.

The information specifying one amino acid is contained in a "codon", a triplet on the mRNA strand recognized by the second type of RNA molecule, the transfer RNA (tRNA). Since any position on the mRNA strand may be occupied by one of four bases, there are $4^3 = 64$ possible codons but only 20 amino acids. Three of the codons are termination signals. The remaining 61 triplets code for amino acids, which suggests the genetic code is degenerate. The amino acids methionine and tryptophan are, for example, each specified by unique codons, while leucine, arginine, and serine may each be specified by one of six codons.[1] Similarly, the other 15 amino acids are each coded for by any of two, three, or four codons.

The third and most abundant type of RNA, comprising about 80% of all cellular RNA, is the ribosomal RNA (rRNA), an important constituent of ribosomes. The precise function of rRNA is not known, although evidence suggests that it is crucial in binding certain components of protein synthesis (such as tRNA and mRNA) to the ribosome.[2]

In 1977, powerful DNA sequencing techniques were announced by Gilbert and Sanger. The concurrent development of these two important DNA sequencing techniques has fostered the rapid accumulation of sequence data; currently, there are about 20 million bases of information available for analysis. A great deal of statistical analysis has been performed on these sequences; we describe some of them briefly.

B. Some Examples of Statistical Studies of DNA

The rapid development of sequencing techniques required the development of computer programs capable of manipulating, analyzing, and comparing long sequences of data. The proliferation of such software is all too obvious: programs are now available which aid in planning cloning projects,[3] predict secondary structure of tRNA and mRNA,[4] determine whether a sequence is protein-coding or noncoding,[5,6] and search for such signals as promoters.[6,7]

Since the bases on DNA encode the information necessary to make each protein, one could suggest that any heterogeneity exceeding random expectation could be a consequence of the difference in the amino acid composition of the various gene products. Elton[8] tested this idea by selecting 42 protein sequences (32 vertebrate and 10 bacterial) with a total combined length of 5801 amino acids. An arbitrary length (20 residues) of each sequence was used to predict the corresponding DNA sequence. Because of the degeneracy of the genetic code, a given protein sequence may be specified by one of several DNA sequences. Elton predicted the base order of each gene fragment, assuming uniform codon preference and then applied analysis of variance to the set of predictions. He concluded that the data (applying uniform codon weighting) were consistent with the suggestion that DNA sequences within genes approximate "random DNA" from the point of view of heterogeneity.[8]

Data, however, suggest that among possible codon choices there are preferences.[9-13] The amino acid glutamine, for example, may be specified by either CAG or CAA, but the CAG codon occurs with much greater frequency in some genes.[14]

Maniatis et al.[14] suggest that knowledge of degenerate codon preference has an important practical application. When screening cellular mRNA for a specific, rare mRNA corresponding to a desired gene product with a partially determined sequence, one can locate the desired mRNA molecule by constructing oligomers complementary to the gene product. Unfortunately, the degeneracy of the genetic code allows even short amino acid sequences to possess a prohibitively large number of possible DNA coding sequences. Taking advantage of codon preference patterns reduces the number of oligomers likely to specify a given protein sequence and can make the mRNA screening process manageable.

Codon preference has been correlated with levels of gene expression in *Escherichia coli* and yeast.[15-18] A strong correlation exists between codon frequency and the relative abundance of the corresponding tRNA.[17,18] Gribskov et al.[16] suggested that codon preferences can be used to predict the relative level of gene expression and give a method to construct preference plots.

Furthermore, codon preferences and other statistical phenomena may be used to determine the function of a given DNA sequence. A common question when analyzing sequence data is whether the bases are part of a coding section (known as an "exon"). As Staden[19] argued, there are two approaches to this problem: one may either infer a strand is proteincoding from such clues as ribosome binding sites and an absence of stop codons or one may examine the properties of the base sequence and determine if they are consistent with properties known to be associated with coding sequences. One such property is codon preference.

Codons are used with unequal frequency in coding sequences. Fickett[20] noted that oligonucleotides (especially single bases) tend to be repeated with a periodicity of three in a protein-coding sequence. Furthermore, such periodicity is absent in noncoding sequences. Coding regions may also be detected using vector Fourier methods.[21] Statistical methods may also be used to determine the reading frame of coding sequence.[22]

Until the development of sophisticated DNA sequencing techniques, only a limited amount of reliable sequence data existed. Among the first nucleic acid data analyzed, therefore, were the base compositions of DNA from various sources. The four nucleotides are not evenly represented in any given sequence, and the base composition varies within and between sequences.[23]

Prokaryotic and eukaryotic DNA sequences also display distinct nearest neighbor patterns. The most basic analyses of nearest neighbor frequencies are the observations of dinucleotide frequencies. Distinct patterns, such as the relative rarity of the CG dinucleotide in eukaryotes[12,24-26] and the preference for PuPu and PyPy pairs* over PuPy and PyPu pairs in eukaryotic DNA[27] have been observed. Nussinov[25-28] has suggested that this preference results from structural considerations, with the homopolymeric dinucleotides (PuPu and PyPy) and other doublets which cause little or no steric strain in the DNA molecule being preferred.

Another example of nucleotide ordering is A clustering in sequences. Nussinov[29] examined long (>1 kb) RNA and DNA sequences for homopolynucleotides. She found that all but one of the sequences (16S rRNA, which does not code for protein, but instead serves a functional role in the ribosome) had fewer single and doubly clustered A than expected. Longer runs of A such as triplets were found more frequently than would be expected from random occurrence. G and C tended to cluster less frequently: single G and C and the doublets GG and CC were present in the analyzed sequences more often than expected, while longer clusters of G and C were observed less frequently than expected. Since a G-C base pair has more hydrogen bonds and, consequently, a higher bond energy than an A-T pair, Nussinov suggested that this clustering may be involved in facilitating the "unzipping" of the DNA strands, expediting replication, transcription, and/or translation.

Nussinov[30,31] has also examined DNA sequences near transcription inititation sites. Evidence suggests that eukaryotic transcription factors recognize and bind to certain promoter regions, including the so-called "CCAAT" and "TATA box" sequences located "upstream"** from the transcriptional start site. Nussinov, however, has found other oligomers arguably as significant as the CCAAT sequence. Eukaryotic sequences within 500 bP of mRNA initiation sites were analyzed for recurring oligomers. TAT/ATA triplets and ATAT/TATA quartets, for example, occur frequently about 275 bases upstream from the start site with a signal strength twice that of the CAAT sequence at -80.[31]

Statistical evaluations have also been performed on single-stranded nucleic acids. The sequences of the single-strand DNA viruses, for example, contain fewer palindromic regions than expected. Palindromes are sequences which, if in double-stranded nucleic acid, would contain a twofold axis of symmetry, e.g., AGCT; consequently, such sequences can fold up on themselves. In one study, Duggleby[32] found improbably few four and six nucleotide palindromes in the single-stranded DNA phage, ΦX174. Other palindromes occurred with a frequency approximately as expected. Duggleby theorized that this paucity of palindromes might be associated with constraints on the secondary structure of single-stranded DNA viruses.

* Pu indicates purine base (A or G), while Py indicates pyrimidine base (C or T).

** "Upstream" suggests base positions in the 5′ direction from the transcription initiation site. The locations of bases upstream are described by negative integers, corresponding to the discrete number line with the start site equal to zero. Similarly, "downstream" implies base positions in the 3′ direction from the transcription start site. The locations of bases downstream are given by positive integers.

III. MARKOV ANALYSIS OF DNA/RNA SEQUENCES

A. Background

The previous section illustrated many examples of the statistical analysis of nucleotide sequences. Many of these are "local" in nature; they use statistics from nearest neighbor frequencies, codon counts, and so on. The very mechanism by which DNA sequences are produced — sequentially in long chains — suggests that the analysis of such sequences might profitably be carried out within the framework of Markov chain methodology. From a statistical point of view, such analyses seem to fall naturally into two camps.

The first might loosely be called "informational analysis". This draws on statistical machinery developed in the late 1950s by Kullback et al.[33] and others. Erickson and Altman[34] used these techniques to search for patterns in the MS2 genome. Rowe and Trainor,[35] Lipman and Maizel,[36] and Lipman and Wilbur[37] use related methods; see also Konopka[38] for a discussion of the evolutionary implications of information content. The second group corresponds to (statistically) more classical Markov chain analysis. See Elton,[24] Almagor,[39] Blaisdell,[40,41] and Garden,[42] for an example.

We first give a brief synopsis of Markov chain terminology. Let $X = \{X_n, n = 0, 1, 2, \ldots\}$ denote a stochastic process whose states represent the nucleotides in a given DNA (or RNA) sequence. For definiteness, we label the bases in alphabetic order, so that A = 1, C = 2, G = 3, and T = 4. X is called a Markov chain of order k if

$$\Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{n-k+1} = i_{n-k+1}, \ldots, X_0 = i_0\}$$
$$= \Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{n-k+1} = i_{n-k+1}\}$$

for all $n \geq k - 1$ and for all choices of states $i_0, i_1, \ldots, i_{n+1}$ from {1,2,3,4}.* Intuitively, this says that the distribution of the next base in the sequence is determined by the previous k bases, and not by earlier ones. When k = 0, the chain comprises independently distributed bases. When k = 1, we recover the usual first order Markov chain case.

The behavior of the (time homogeneous) process X is determined by its initial distribution and the transition probabilities $p(i_1, i_2, \ldots, i_k; i_{k+1})$ given by

$$p(i_1, i_2, \ldots, i_k; i_{k+1}) = \Pr\{X_{k+1} = i_{k+1} \mid X_k = i_k, \ldots, X_1 = i_1\} \quad (1)$$

The aim is to estimate the order k of the model, the transition probabilities $p(i_1, i_2, \ldots, i_k; i_{k+1})$, and then to assess various hypotheses about the DNA sequence(s). Typical among analyses for a single sequence might be testing for independence (k = 0), testing for uniformity of base composition, testing a fit to a hypothesized transition matrix, and testing particular transition probabilities. For a collection of sequences (perhaps derived as subsequences of a given sequence), one is usually interested in finding heterogeneity among the sequences. Within the framework of first order Markov chains, these questions are addressed by Elton.[24] Similar questions in higher dimensions may be studied using standard theory; References 43 and 44 are recommended.

Rather than focus on such detailed questions, our interest will focus on statistical aspects of finding the order of the Markov chain.

B. Finding the Order of a Markov Chain

Suppose that the sequence of interest is of length N. For r = 1,2, . . . let $n(i_1, i_2, \ldots, i_r)$

* There may be reasons to consider alphabets other than {A,C,G,T}, for example, {Pu,Py}. For simplicity, we stay with the "nucleotide alphabet". The methods to be described later carry over, with obvious changes, to other choices.

be the number of transitions $i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow i_r$ observed in the sequence. It is a standard result[43,44] that the maximum likelihood estimator of $p(i_1,i_2, \ldots ,i_k; i_{k+1})$ is

$$\hat{p}(i_1,i_2, \ldots, i_k; \quad i_{k+1}) = n(i_1, i_2, \ldots, i_k, i_{k+1})/n(i_1, i_2, \ldots, i_k, +) \tag{2}$$

where

$$n(i_1, i_2, \ldots,i_k, +) = \sum_j n(i_1,i_2, \ldots, i_k, j)$$

Notice that in fitting a k-th order Markov chain with m states, there are $p = m^k(m - 1)$ independent parameters to be estimated; in our case $m = 4$, so that

$$p = 3 \times 4^k \tag{3}$$

This gives $p = 3$ ($k = 0$), $p = 12$ ($k = 1$), $p = 48$ ($k = 2$), $p = 192$ ($k = 3$), $p = 768$ ($k = 4$), and $p = 3072$ ($k = 5$). It is clear from this that very long DNA sequences are required for "good" estimation of the transition probabilities under the fully parameterized model of Equation 2. We will return to this problem later.

There have been several methods proposed for estimating the order k of a Markov chain. Because we will later want to compare models which are not nested, we will use information criteria rather than a multiple hypothesis testing framework. Among these is a standard information theory method,[43] Akaike's information criterion (AIC),[42,45,46] and the Bayesian information criterion (BIC).[45,47] To compute the BIC, we evaluate the log-likelihood L of the data

$$L = \sum n(i_1,i_2, \ldots, i_k, i_{k+1}) \ell n \hat{p}(i_1,i_2, \ldots, i_k; \quad i_{k+1}) \tag{4}$$

the sum being over all $i_1,i_2, \ldots ,i_k,i_{k+1}$ for which $n(i_1,i_2, \ldots ,i_k,i_{k+1})>0$. The BIC for order k is then defined by

$$BIC(k) = -2 L + p \ln n \tag{5}$$

where n ($\leq N$) is the number of subsequences from which the counts $n(\cdot)$ were formed. That k which minimizes BIC(k) is taken as the estimator of the true order of the chain. We use BIC because (unlike AIC) it is a consistent estimator of Markov chain order,[45] and it chooses simpler models.

As noted above, a full Markov chain analysis of high order requires very long data sequences. Because of the inherent heterogeneity of the linear structure of DNA sequences, such long *homogeneous* sequences are rather unusual. The data prevent us from performing precisely the type of analysis that seems most interesting. A class of Markov chain models that combines high order dependence with a small number of parameters could prove useful.

C. Models for High-Order Markov Chains

The class of Markovian models we will use in the present analysis was developed by Raftery.[47] The typical transition probability of the k-th order model is of the form

$$p(i_1,i_2, \ldots, i_k; \quad i_{k+1}) = \sum_{j=1}^{k} \lambda_j q(i_j,i_{k+1}) \tag{6}$$

where $Q = \{q(i,j), 1 \leq i,j \leq 4\}$ is a row-stochastic matrix whose entries are to be estimated

Table 1
FIVE BIOLOGICALLY INTERESTING REGIONS

Sequence	Positions	Orientation*	Length, N
Late	45000-19600	L→R	23102
Early 2	38220-45000	L→R	6780
Early 1	35040-27581	R→L	7460
Control	38030-33290	R→L	4740
Silent	27580-19601	R→L	7980

* If orientation is R→L, sequence is read in reverse complement form.

from the data, and $\lambda_1,\lambda_2, \ldots ,\lambda_k$ are k parameters, summing to 1, that must also be estimated. This model may be viewed as a discrete state-space analog of the auto-regressive time series models; one extra parameter is introduced for each extra order after the first.

Notice that when $k = 1$, this model is identical to the usual first order Markov chain described earlier. The number of parameters to be estimated in the k-th order case is reduced from Equation 3 to

$$p = 11 + k \tag{7}$$

and so we should be able to look for high order dependence more successfully. The price we pay for this is that the algebraic simplicity of Equation 2 no longer applies, and the maximum likelihood estimates of the parameters must be found by numerically maximizing the log-likelihood

$$L = \sum n(i_1,i_2, \ldots, i_k, i_{k+1}) \ell n \left(\sum_{j=1}^{k} \lambda_j q(i_j,i_{k+1})\right) \tag{8}$$

using a constrained nonlinear optimization algorithm.

The BIC for this model is computed using Equation 5; L is the value of the right-hand side of Equation 8 at the maximum likelihood estimates.

D. Examples

We have chosen the bacteriophage λ as the source of our example. The λ genome was sequenced by Sanger et al.[48] and is 48502 nucleotides in length. The sequence was initially broken into five biolgocially interesting regions, given in Table 1.

The BIC indexes given by Equation 5 were calaculated for each sequence. For comparative purposes, we also include a value of $k = -1$ which corresponds to the model of independent bases, each with relative frequency 1/4. In this case, there are no parameters to estimate, and the BIC value is

$$BIC(-1) = -2n \ell n 4$$

where n is the effective number of observations in the data. In all the results presented here, the sequence comparisons were begun at the 7th base of each sequence, so that $n = N - 6$ may be calculated from Table 1. The results of this analysis are presented in Table 2.

The BIC criterion indicates first order dependence for the regions Silent, Early 1, Early 2, and Control. We might expect the Silent region to have less "structure" than the others,

Table 2
BIC(k) VALUES FOR BACTERIOPHAGE λ

Sequence	Late	Early 1	Early 2	Control	Silent
k = −1	64036	20667	18782	13125	22109
0	63682	20566	18682	13025	21971
1	63258	20472*	18618*	12982*	21900*
2	62349*	20648	18743	13218	22026
3	63158	21706	19771	14133	23041
4	67923	26149	24040	18345	27580

* Denotes order of chain using the BIC criterion.

Table 3
LOCATION OF LATE REGIONS

Sequence	Positions	Orientation*	Length, N
Head	1-8550	L→R	8550
Tail	8550-19600	L→R	11050
Lysis	45000-46430	L→R	1430

* If orientation is R→L, sequence is read in reverse complement form.

since λ can propagate without this region.[14] The Late region is clearly identified as having an order of dependence of 2.

Models which are so clearly described by first order chains as those here will not typically be improved by using the high order dependence (HOD) models of Equation 6. We therefore analyzed only the Late region to determine whether the HOD model gives a more parsimonious description of the data. The BIC values are 62927 (k = 2) and 62935 (k = 3), compared to the smallest value of 62349 (k = 2) for the general case (see Table 2). The HOD model of order 2 provides the second best description of the data. That it is not the best means that the second order transition matrix of the best model must have a more complicated structure than is contained in Equation 6. The HOD models should provide better models for sequences such as φX174[42] that exhibit a higher order of dependence than λ.

The low orders of dependence identified in these regions may be due in part to inhomogeneity in the sequences. To examine this further, we broke the late region into three subregions labeled Head, Tail, and Lysis. The locations of these regions are given in Table 3 and the corresponding BIC values in Table 4.

The overall appearance of a second order model is maintained. Remarkably enough, the Lysis region is adequately described by the "completely random" model in which bases are laid down uniformly and at random. The Lysis region is composed of three genes, S, R, and RZ.[48] R and S genes are necessary for lysing the bacterium after the production of progeny phage and so are essential to the propagation of the phage in nature. Naively perhaps, we expected the structure of these genes to be similar to other λ genes (i.e., showing some dependence). Individual gene sequences, however, show the same completely random structure as the entire Lysis region. We conjecture that the variability in the estimated orders of dependence of these coding regions can be attributed to different patterns of codon usage among the genes.

We also examined one of the open reading frames in the Early 2 region, ORF290. Once again, a model of independent bases is adequate. It is tempting to conclude that ORF290

Table 4
BIC(k) VALUES FOR BACTERIOPHAGE λ

Sequence	Head	Tail	Lysis
k = −1	23689	30620	3948*
0	23532	30287	3966
1	23352	30152	3997
2	23113*	29670*	4193
3	24114	30567	5017
4	28439	35070	—

* Denotes order of chain using the BIC criterion.

has no coding function, but our experience in the Lysis region (and that of Garden[42] for the replicase gene of MS2) demonstrates that protein coding regions may often appear structureless.

IV. TRANSFORM ANALYSIS OF DNA/RNA SEQUENCES

A. Background

The previous section of this article is devoted to stochastic models for the analysis of the primary structure of one or more stretches of DNA. These methods are useful for finding parsimonious descriptions of stretches of sequence with little *apparent* structure. One interesting feature of eukaryotic DNA is the presence of tandem (or periodic) repeats and interspersed base sequence repeats throughout the genome. Such repeats vary in length from simple dinucleotide periodicities (for example, the dinucleotide AG [repeated 28 times] found upstream from the mouse immunoglobulin G3 constant region gene[49]) to very large tandem repeats; the GNOMIC dictionary[50] is an invaluable compilation of examples of this type. Nucleotide sequences that exhibit repetitive structure cannot usefully be described by the earlier Markov chain models; alternatives are needed.

The presence of periodicities in DNA (or protein) sequences has led several authors to use what might broadly be called Fourier analysis methods to search for such structure. Kubota et al.[51] and McLachlan and Karn[52] use correlation coefficients calculated from sequences when residues are replaced by various quantitative properties of the amino acids, such as hydrophobicity. Liquori and co-workers[21] have introduced several Fourier-analytic methods for studying sequence similarities between proteins of different species. Felsenstein et al.[53] suggested Fourier analysis as a fast technique for computing the fraction of matches between two large nucleic acid sequences. Silverman and Linsker[54] and Trifonov and Sussman[55] use related methods to detect regularities in DNA sequences.

In this section we describe another technique of sequency analysis based on the Walsh transform, that is also applicable to problems involving symbol-sequence periodicities.

B. The Walsh Transform

We first give an inductive definition of the Walsh functions $\{W_n(x), 0 \le x < 1\}$, $n = 0,1, \ldots$. We initialize the induction by defining

$$W_0(x) = 1, \quad x \in [0,1)$$

$$W_1(x) = \begin{cases} 1, & x \in [0,1/2) \\ -1, & x \in [1/2,1) \end{cases} \tag{9}$$

and then proceed recursively for n = 1,2, . . . via

$$W_{2n}(x) = \begin{cases} W_n(2x), & 0 \leq x < 1/2 \\ (-1)^n W_n(2x-1), & 1/2 \leq x < 1 \end{cases} \tag{10}$$

and

$$W_{2n+1}(x) = \begin{cases} W_n(2x), & 0 \leq x < 1/2 \\ (-1)^{n+1} W_n(2x-1), & 1/2 \leq x < 1 \end{cases} \tag{11}$$

The Walsh functions are orthogonal and piecewise constant on [0,1).[56] The even numbered Walsh functions are symmetric about $x = {}^1/_2$ (and play the role of the cosine terms in Fourier series), while the odd-numbered Walsh functions are antisymmetric about $x = {}^1/_2$ (and so play the role of the sine terms in Fourier series). These functions may be used to construct the discrete Walsh transform of the sequence $\mathbf{x} = (x_0, x_1, \ldots, x_{N-1})$. When N is a power of 2 (so that $N = 2^p$ for some positive integer p), the following recipe does the trick.

We define first

$$w(k,j) = W_k(j/N), \quad j = 0,1, \ldots, N-1 \tag{12}$$

The Walsh transform of x is then given by

$$a_k = \frac{1}{N} \sum_{j=0}^{N-1} x_j \, w(k,j), \quad k = 0, 1, \ldots, N-1 \tag{13}$$

and the inverse transform[56] by

$$x_j = \sum_{k=0}^{N-1} a_k w(j,k), \quad j = 0,1, \ldots, N-1 \tag{14}$$

There is an explicit formula for w(k,j).[56] If $j = \sum_{r=0}^{p-1} j_r 2^r$, and $k = \sum_{r=0}^{p-1} k_r 2^r$, then

$$w(k,j) = (-1)^{\sum_{r=0}^{p-1} j_r (k_{p-r} + k_{p-r-1})} \tag{15}$$

from which it readily follows that w(k,j) = w(j,k).

We will use the discrete Walsh transform to hunt for periodicities in DNA sequences. Imagine that our sequence of length $N = 2^p$ is listed as $A_0, A_1, \ldots, A_{N-1}$, where (as in Section III) A_i takes the value 1 if the i-th base is an A;2, if it is a C; 3, if it is G, and 4 if it is T. We generate four associated sequences $\{x_{ij}, j = 0,1, \ldots, N-1\}$ of indicators as follows. For $i = 1, \ldots, 4$, set

$$x_{ij} = \begin{cases} 1, & \text{if } A_j = i, \quad j = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Then form the associated Walsh transforms using Equation 13:

$$a_{ik} = \frac{1}{N}\sum_{j=0}^{N-1} x_{ij}w(k,j), \quad k = 0,1, \ldots, N-1 \tag{17}$$

Notice that a_{i0} is the fraction of base i in the sequence.

The power spectrum $\{c_k, k = 0,1, \ldots, N-1\}$ of the sequence x is then defined by

$$c_k = \sum_{i=1}^{4} a_{ik}^2 \tag{18}$$

Notice that from a computational point of view, only three of the transform sequences $\{a_{ik}, k = 0,1, \ldots, N-1\}$ need to be calculated. Since

$$\sum_{i=1}^{4} x_{ij} = 1, \quad j = 0,1, \ldots, N-1$$

it follows that

$$\begin{aligned}\sum_{i=1}^{4} a_{ik} &= \sum_{i=1}^{4}\left(\frac{1}{N}\sum_{j=0}^{N-1} x_{ij}w(k,j)\right)\\ &= \frac{1}{N}\sum_{j=0}^{N-1}\left(\sum_{i=1}^{4} x_{ij}\right)w(k,j)\\ &= \frac{1}{N}\sum_{j=0}^{N-1} w(0,j)w(k,j)\\ &= \begin{cases}1, & \text{if } k = 0\\ 0, & \text{if } k \neq 0\end{cases}\end{aligned}$$

the last equality following from orthogonality.

We will now compare the properties of the Walsh power spectrum with the perhaps more familiar Fourier power spectrum, found by replacing Equation 17 with

$$a_{mk} = \frac{1}{N}\sum_{j=0}^{N-1} x_{mj}\, e^{2\pi ijk/N}, \quad k = 0,1, \ldots, N-1 \tag{19}$$

and the value of c_k in Equation 18 by

$$c_k = \sum_{i=1}^{4} |a_{ik}|^2 \tag{20}$$

There are alternative ways of representing the structure of a DNA sequence other than via the indicator variables used in Equation 16. Silverman and Linsker,[54] for example, use a tetrahedral coordinate representation. The computational algorithm used here to calculate the fast Fourier transform is based on the code of Press et al.[57]

C. Examples

Figures 1 through 3 display the spectra for a 128 bP consensus sequence from the AT-

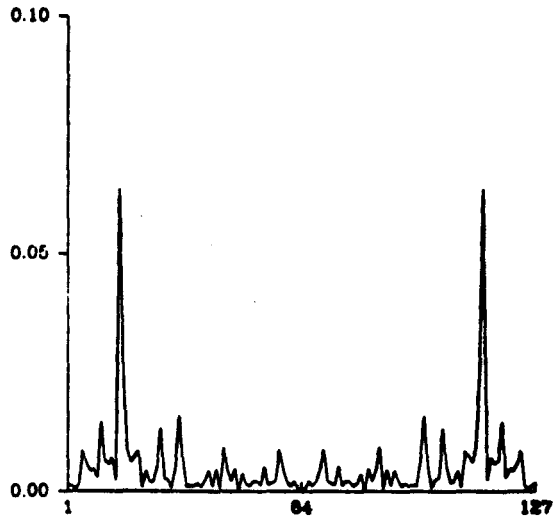

FIGURE 1. Fourier power spectrum for 128 base sequence from *X. laevis* oocyte 5S DNA.

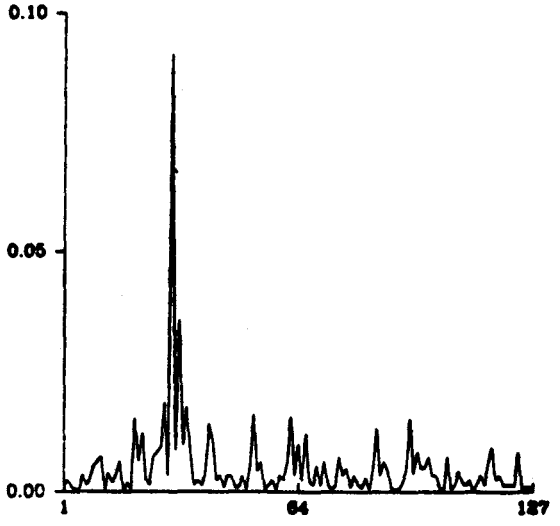

FIGURE 2. Walsh power spectrum for 128 base sequence from *X. laevis* oocyte 5S DNA.

rich spacer region of the 5S DNA of *Xenopus laevis*.[58] Figure 1 illustrates the familiar symmetry of the Fourier power spectrum. The pronounced peak at $k = 16$ shows a periodicity of length 8, corresponding to the simple sequence repeat 8 bases in length.[54,58] In contrast, the Walsh spectrum (Figure 2) is not symmetric; now the peak near $k = 32$ is indicative of a periodic component of length 8. Figure 3 gives the graphs of both transforms, plotted on common axes for comparative purposes. Note that in these plots (and those that follow), the base-frequency information that corresponds to $k = 0$ is not plotted.

The second example is a 128-bP sequence from the human ξ-globin gene[59]; it starts at position 210, in Intron 1. There is a 14 base repeat sequence ACAGTGGGAGGGG repeating (with very little variation) through this region. Notice from Figures 4 and 5 that both power spectra are considerably less well defined, despite the presence of the repeat. There are

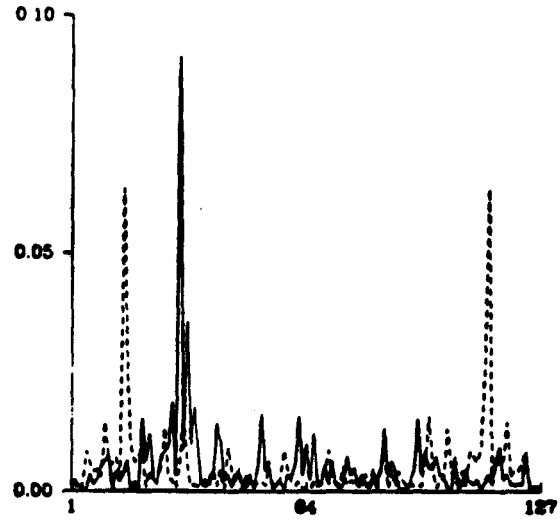

FIGURE 3. Composite of Figures 1 and 2. Dotted lines correspond to Fourier transform and solid lines to the Walsh transform.

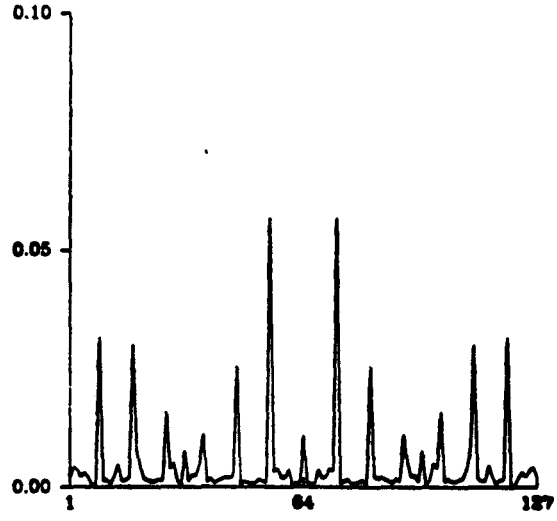

FIGURE 4. Fourier power spectrum for 128 base sequence from human ξ-globin, intron 1.

several reasons for this, among them, the rather repetitive substructure of the repeat and the fact that maximum emphasis of the peaks will occur when the length of the sequence is a multiple of the length of the repeat. Figure 6 gives the superposition of the two graphs.

There are several other comments worth making about use of the discrete Walsh transform in this setting. First, the technique is not limited to the analysis of sequences which have length a power of 2, however, when the length is a power of 2, the computation of the transform coefficients is very simple, both in terms of computational code and speed of execution. The fast Walsh transform requires less storage space than the corresponding fast Fourier transform, and seems to execute about five times faster. On the other hand the Walsh transform is, by its nature, better adapted to hunting for periodicities that are powers of 2

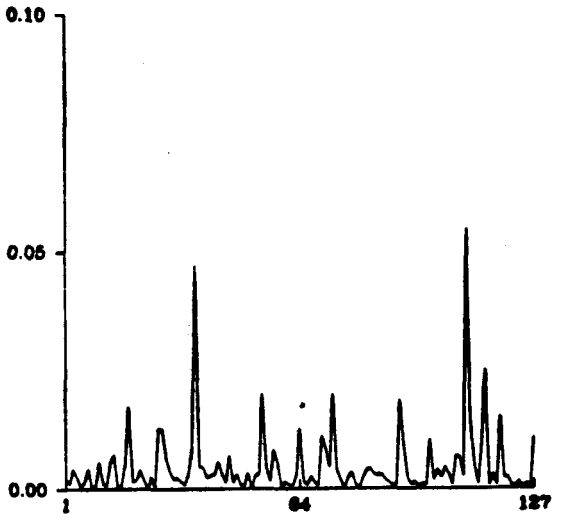

FIGURE 5. Walsh power spectrum for 128 base sequence from human ξ-globin, intron 1.

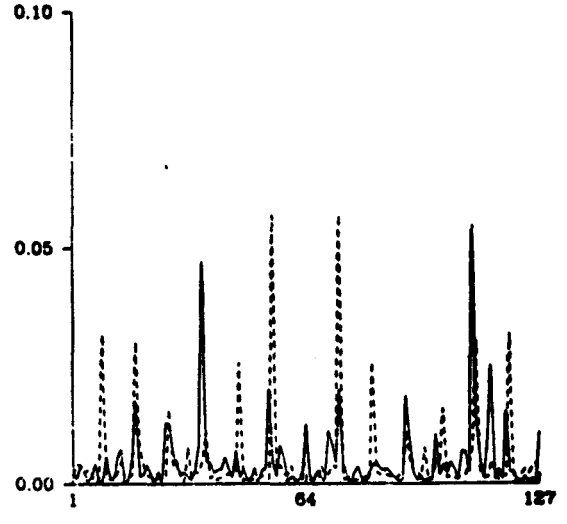

FIGURE 6. Composite of Figures 4 and 5. Dotted lines correspond to Fourier transform and solid lines to the Walsh transform.

in length, and it is not always easy to interpret the form of the power spectrum. Nevertheless, both techniques are a useful augmentation to more traditional oligonucleotide dictionary searches in the hunt for symbol patterns.

ACKNOWLEDGMENTS

This paper forms part of Bart Giddings' Senior Honors thesis at the University of Utah.

Work by Simon Tavaré is supported in part by NSF grant DMS86-08857 and in part by a grant from the System Development Foundation to Michael Waterman.

We would like to thank Jim Keener, Adrian Raftery, and John Roth for useful discussions on some computational and biochemical aspects of this work.

REFERENCES

1. **Freifelder, D.,** *Essentials of Molecular Biology*, Jones & Bartlett Publishers, 1985.
2. **Suzuki, D. T., Griffiths, A. J. F., Miller, J. H., and Lewontin, R. C.,** *An Introduction of Genetic Analysis*, 3rd ed., W. H. Freeman, New York, 1986.
3. **Blumenthal, R. M., Rice, P. J., and Roberts R. J.,** Computer programs for nucleic acid sequence manipulation, *Nucl. Acids Res.*, 10, 91, 1982.
4. **Nussinov, R. and Pieczenik, G.,** Structural combinatorial constraints on base pairing in large nucleotide sequences, *J. Theor. Biol.*, 106, 245, 1984.
5. **Tramontano, A. and Macchiato, M. F.,** Probability of coding a DNA sequence: an algorithm to predict translated reading frames from their thermodynamic characteristics, *Nucl. Acids Res.*, 14, 127, 1986.
6. **Staden, R.,** Graphic methods to determine the function of nucleic acid sequences, *Nucl. Acids Res.*, 12, 521, 1984.
7. **Mulligan, M. E. and McClure, W. R.,** Analysis of the occurrence of promoter-sites in DNA, *Nucl. Acids Res.*, 14, 109, 1986.
8. **Elton, R. A.,** Theoretical models for heterogeneity of base composition in DNA, *J. Theor. Biol.*, 45, 533, 1974.
9. **Miyata, T., Hayashida, H., Yasunaga, T., and Hasegawa, M.,** The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other, *Nucl. Acids Res.*, 7, 2431, 1979.
10. **Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A.,** Codon catalog usage and the genome hypothesis, *Nucl. Acids Res.*, 8, r49, 1980.
11. **Grantham, R., Gautier, C., and Gouy, M.,** Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type, *Nucl. Acids Res.*, 8, 1893, 1980.
12. **Rothberg, P. G. and Wimmer, E.,** Mononucleotide and dinucleotide frequencies and codon usage in poliovirion RNA, *Nucl. Acids Res.*, 9, 6221, 1981.
13. **Modiano, G., Battistuzzi, G., and Motulsky, A. G.,** Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects?, *Proc. Natl. Acad. Sci. U.S.A.*, 78, 1110, 1981.
14. **Maniatis, T., Fritsch, E. F., and Sambrook, J.,** *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1982.
15. **McLachlan, A. D., Staden, R., and Boswell, D. R.,** A method for measuring the non-random bias of a codon usage table, *Nucl. Acids Res.*, 12, 9567, 1984.
16. **Gribskov, M., Devereux, J., and Burgess, R. R.,** The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression, *Nucl. Acids Res.*, 12, 539, 1984.
17. **Ikemura, T.,** Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.*, 146, 1, 1981.
18. **Ikemura, T.,** Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes, *J. Mol. Biol.*, 158, 573, 1982.
19. **Staden, R.,** Measurement of the effects that coding for a protein has on a DNA sequence and their use for finding genes, *Nucl. Acids Res.* 12, 551, 1984.
20. **Fickett, J. W.,** Recognition of protein coding regions in DNA sequences, *Nucl. Acids Res.*, 10, 5303, 1982.
21. **Liquori, A. M., Ripamonti, A., Sadun, C., Ottani, S., and Braga, D.,** Pattern recognition of sequence similarities in globular proteins by Fourier analysis: a novel approach to molecular evolution, *J. Mol. Evol.*, 23, 80, 1986.
22. **Shulman, M. J., Steinberg, C. M., and Westmoreland, N.,** The coding function of nucleotide sequences can be discerned by statistical analysis, *J. Theor. Biol.*, 88, 409, 1981.
23. **Weir, B. S.,** Statistical analysis of molecular genetic data, *IMA J. Math. Appl. Med. Biol.*, 2, 1, 1985.
24. **Elton, R. A.,** Doublet frequencies in sequenced nucleic acids, *J. Mol. Evol.*, 4, 323, 1975.
25. **Nussinov, R.** Some rules in the ordering of nucleotides in the DNA, *Nucl. Acids Res.*, 8, 4545, 1980.
26. **Nussinov, R.,** The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice, *J. Mol. Evol.*, 17, 237, 1981.
27. **Nussinov, R.,** Strong doublet preferences in nucleotide sequences and DNA geometry, *J. Mol. Evol.*, 20, 111, 1984.
28. **Nussinov, R.,** Doublet frequencies in evolutionary distinct groups, *Nucl. Acids Res.*, 12, 1749, 1984.
29. **Nussinsov, R.,** Strong adenine clustering in nucleotide sequences, *J. Theor. Biol.*, 85, 285, 1980.
30. **Nussinov, R.,** Compilation of eukaryotic sequences around transcription initiation sites, *J. Theor. Biol.*, 120, 479, 1986.
31. **Nussinov, R., Owens, J., and Maizel, J. V., Jr.,** Sequence signals in eukaryotic upstream regions, *Biochim. Biophys. Acta*, 866, 109, 1986.

32. **Duggleby, R. G.**, A paucity of palindromes in ϕX174, *J. Theor. Biol.*, 93, 143, 1981.
33. **Kullback, S., Kupperman, M., and Ku, H. H.**, Tests for contingency tables and Markov chains, *Technometrics*, 4, 573, 1962.
34. **Erickson, J. W. and Altman, G. G.**, A search for patterns in the nucleotide sequence of the MS2 genome, *J. Math. Biol.*, 7, 219, 1979.
35. **Rowe, G. W. and Trainor, L. E. H.**, On the informational content of viral DNA, *J. Theor. Biol.*, 101, 151, 1983.
36. **Lipman, D. J. and Maizel, J.**, Comparative analysis of nucleic acid sequences by their general constraints, *Nucl. Acids Res.*, 10, 2723, 1982.
37. **Lipman, D. J. and Wilbur, W. J.**, Contextual constraints on synonymous codon choice, *J. Mol. Biol.*, 163, 363, 1983.
38. **Konopka, A.**, Is the information content of DNA evolutionarily significant?, *J. Theor. Biol.*, 107, 697, 1984.
39. **Almagor, H.**, A Markov analysis of DNA sequences, *J. Theor. Biol.* 104, 633, 1983.
40. **Blaisdell, B. E.**, A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences, *J. Mol. Evol.*, 19, 122, 1983.
41. **Blaisdell, B. E.**, Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding, *J. Mol. Evol.*, 21, 278, 1985.
42. **Garden, P. W.**, Markov analysis of viral DNA/RNA sequences, *J. Theor. Biol.* 82, 679, 1980.
43. **Chatfield, C.**, Statistical inference regarding Markov chain models, *Appl. Stat.*, 22, 7, 1973.
44. **Basawa, I. V. and Prakasa Rao, B. L. S.**, *Statistical Inference for Stochastic Processes*, Academic Press, New York, 1980.
45. **Katz, R. W.**, On some criteria for estimating the order of a Markov chain, *Technometrics*, 23, 243, 1981.
46. **Tong, H.**, Determination of the order of a Markov chain by Akaike's Information Criterion, *J. Appl. Probab.*, 12, 488, 1975.
47. **Raftery, A. E.**, A model for high-order Markov chains, *J. R. Stat. Soc. Ser. B*, 47, 528, 1985.
48. **Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Petersen, G. B.**, Nucleotide sequence of bacteriophage Lambda DNA, *J. Mol. Biol.*, 162, 729, 1982.
49. **Weis, J. A., Word, C. J., Rimm, D., Der-Balan, G. P., Martinez, H. M., Tucker, P. W., and Blattner, F. R.**, Structural analysis of the murine IgG3 constant region gene, *EMBO J.*, 3, 2041, 1984.
50. **Trifonov, E. N. and Brendel, V.**, *GNOMIC. A Dictionary of Genetic Codes*, Balaban Publishers, 1986.
51. **Kubota, Y., Takahashi, S., Nishikawa, K., and Ooi, T.**, Homology in protein sequences expressed by correlation coefficients, *J. Theor. Biol.*, 91, 347, 1981.
52. **McLachlan, A. D. and Karn, J.**, Periodic features of the amino acid sequence of the nematode myosin rod, *J. Mol. Biol.*, 164, 605, 1983.
53. **Felsenstein, J., Sawyer, S., and Kochin, R.**, An efficient method for matching nucleic acid sequences, *Nucleic Acids Res.*, 10, 133, 1982.
54. **Silverman, B. D. and Linsker, R.**, A measure of DNA periodicity, *J. Theor. Biol.*, 118, 295, 1986.
55. **Trifonov, E. N. and Sussman, J. L.**, The pitch of chromatin DNA is reflected in its nucleotide sequence, *Proc. Natl. Acad. Sci. U.S.A.*, 77, 3816, 1980.
56. **Kennett, B. L. N.**, Introduction to the Finite Walsh Transform and the theory of the Fast Walsh Transform, in Proc. Conf. Theory and Application of Walsh Functions, Hatfield Polytechnic, London, 1971.
57. **Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.**, *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, London, 1986.
58. **Federoff, N. V. and Brown, D. D.**, The nucleotide sequence of oocyte S DNA in *Xenopus laevis*. I. The AT-rich spacer, *Cell*, 13, 701, 1978.
59. **Proudfoot, N. J., Gil, A., and Maniatis, T.**, The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene, *Cell*, 31, 553, 1982.