An optimal DNA segmentation based on the MDL principle

Wojciech Szpankowski* and Wenhui Ren

Department of Computer Science,
Purdue University,
West Lafayette, IN 47907, USA
E-mail: spa@cs.purdue.edu E-mail: renwh@cs.purdue.edu
*Corresponding author

Lukasz Szpankowski

Cell and Molecular Biology,
University of Michigan,
Ann Arbor, MI 48104, USA
E-mail: lszpanko@umich.edu

Abstract: The biological world is highly *stochastic* and *inhomogeneous* in its behaviour. There are regions in DNA with high concentration of G or C bases; stretches of sequences with an abundance of CG dinucleotide (CpG islands); coding regions with strong periodicity-of-three pattern, and so forth. Transitions between these regions of DNA, known also as change points, carry important biological information. Computational methods used to identify these homogeneous regions are called *segmentations*. Viewing a DNA sequence as a *non-stationary* process, we apply recent novel techniques of universal source coding to discover stationary (possibly recurrent) segments. In particular, the Stein-Ziv lemma is adopted to find an asymptotically optimal discriminant function that determines whether two DNA segments are generated by the same source assuring exponentially small false positives. Next, we use the Minimum Description Length (MDL) principle to select parameters that leads to a linear-time segmentation algorithm. We apply our algorithm to human *chromosome* 9 and *chromosome* 20 to discover coding and noncoding regions, starting positions of genes, as well as the beginning of CpG islands.

Keywords: DNA segmentation; stochastic modelling; universal data compression; MDL principle; model selection; piecewise stationary sequences.

Reference to this paper should be made as follows: Szpankowski, W., Ren, W-H. and Szpankowski, L. (2005) 'An optimal DNA segmentation based on the MDL principle', *Int. J. Bioinformatics Research and Applications*, Vol. 1, No. 1, pp.3–17.

Biographical notes: Wojciech Szpankowski received his MS and PhD degrees in Electrical and Computer Engineering from the Technical University of Gdansk in 1976 and 1980, respectively. Currently, he is Professor of Computer Science and by courtesy of Electrical and Computer Engineering at the Purdue University. Szpankowski's research interests cover analysis of algorithms, information theory, bioinformatics, analytic combinatorics, stability problems, and applied probability. He has published over 150 papers on these topics, and a book *Average Case Analysis of Algorithms on Sequences*. He is on the

editorial boards of *Theoretical Computer Science, ACM Trans. on Algorithms, Foundation and Trends in Communications* and *Information Theory, and Discrete Mathematics and Theoretical Computer Science*. He serves on the Steering Committee of the Analysis of Algorithms.

Wenhui Ren received his BS and MS degrees in Computer Science from the Purdue University in 2002 and 2004, respectively. Since 2002, he has been working as a graduate Research Assistant on bioinfomatic projects sponsored by the NSF and NIH. His research interests cover bioinformatics and application of information theory to computational biology.

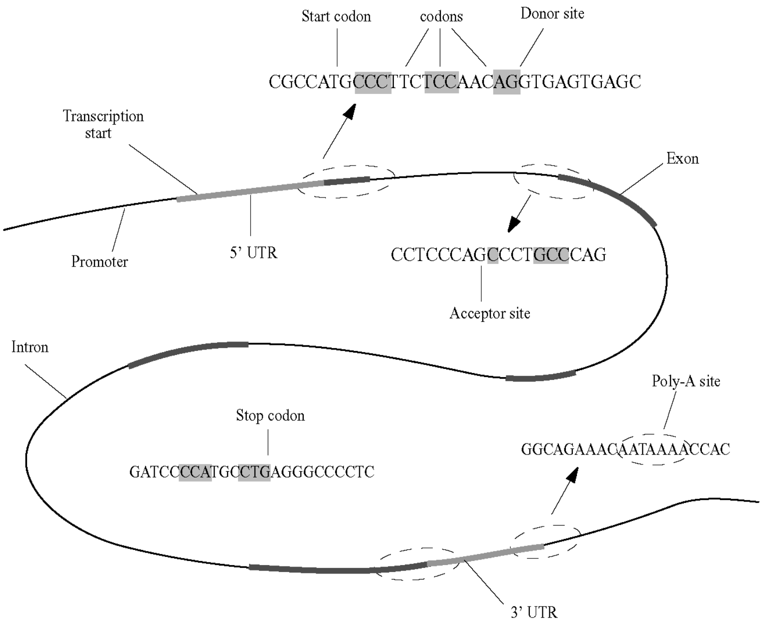
Lukasz Szpankowski is currently an undergraduate senior at the University of Michigan, and is expecting to graduate with a BS in Cell and Molecular Biology in the spring of 2005. His interests include molecular biology, bioinformatics, and preventive medicine.

1 Introduction

This work is built on two premises: On the one hand, a major challenge facing computational biology is the post-sequencing analysis of genomic DNA sequences, and in general biological sequences. It is known that a typical DNA sequence is not homogeneous and that (approximate) repetitions as well as regions with varying statistical properties may have biological meanings (e.g., regulatory elements, structural features of the DNA, CpG islands, coding vs. noncoding regions (Pevzner, 2000; Salzberg et al., 1999)). Computational methods used to identify these homogeneous regions are called segmentations. On the other hand, information theory (Cover and Thomas, 1991; Szpankowski, 2001), in particular source coding better known as data compression, explored patterns, repetitions, motifs, and sequence complexity to design optimal and universal compression schemes, for sequences generated by unknown sources. In *universal* data compression schemes the first (and the most important) step is the *modelling* part that finds a fit between data and a statistical model. This step is also common to many pattern discovery algorithms.

For the highly stochastic and inhomogeneous biological world (Salzberg et al., 1999), probabilistic methods such as the hidden Markov model, Bayesian approach, large deviations, combinatorial tools based on graphs, enumeration methods, and combinatorial optimisations are all well recognised techniques of computational biology (Baldi and Brunak, 2000; Clote and Backofen, 2000; Pevzner, 2000). In this paper, we apply rigorous methods of information theory to DNA segmentation and find recurrent sources in a DNA sequence (Bernaola-Galván et al., 2000; Li, 2001a, 2001b; Sahinalp et al., 2002). In particular, we use techniques developed for non-stationary sequences (Shamir and Merhav, 1999; Shamir and Costello, 2000). Furthermore, following recent experimental results (Nevill-Manning and Witten, 1999; Tabus et al., 2003), we adopt here the view that biological sequences are optimally compressed by nature. We then reverse this process, keeping optimal setting in place, to find structural properties of such sequences. This allows us to construct a linear-time segmentation algorithm.

Figure 1 The structure of a DNA with some additional signals shown



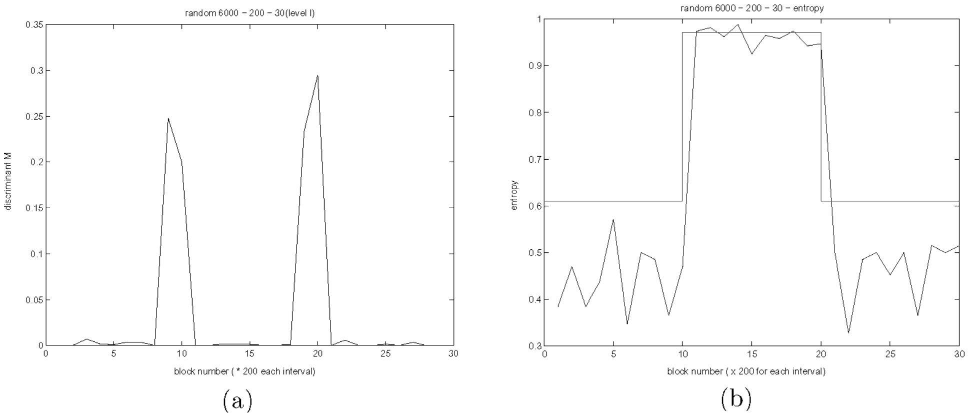
To justify our approach from a biological point of view, we briefly review some basic biological facts and present examples of non-stationary nature of biological sequences. We refer to Figure 1 (Salzberg et al., 1999).

- Oligomer frequency can be used to distinguish between coding and noncoding regions (i.e., 'junk DNA' also known as 'inter-gene region'). We observe that coding regions are rather short (averaging 150 nucleotides in length). It is known that coding regions are translated in triples while other regions are not. Therefore, the codon frequency of these regions might provide a distinctive statistical signature.
- Human gene mutations have been observed more frequently at certain dinucleotides than others. For example, in the dinucleotide CpG, the C nucleotide is often chemically modified by methylation leading to CpG being rarer in the genome than would be expected from the independent probabilities of C and G. Actually, for biological reasons, the methylation process is suppressed in short stretches of the genome, for example, around the promoters or start regions. As a consequence, one observes the so-called CpG islands of typically a few hundred to a few thousands bases long.
- Short oligomers can be used to detect a variety of important biological features in a sequence. For example, the region just before the splice site at the end of an intron typically has high C + T content. This is detectable by statistical methods, but the signal is too weak and often many false positives are detected.
- The ribosome binding site, which occurs just before the start codon initiates protein translation, shows a clear preferences for some bases over others at particular positions. In addition, there is a strong correlation between adjacent positions in the site.

Segmentation, also known as change point analysis or partitioning/fragmentation, was studied in various fields ranging from image processing to statistics. Earlier work on DNA segmentation are due to Churchill (1989), Crowley et al. (1997) (hidden Markov model), and others (e.g., moving window approach, walking Markov model). Techniques used range from Bayesian approach (Ramensky et al., 2000; Sahinalp et al., 2002), model selection (Li, 2001a, 2001b), and entropy-related techniques (Bernaola-Galván et al., 1996, 1999, 2000). One particular attractive segmentation is the divide-and-conquer approach (Bernaola-Galván, 1996) in which DNA is first segmented into two substrings for which the base composition is maximised, and this is carried on the left and right substrings. However, all the above techniques are computationally involved, and there are no clear criteria for stopping the recursion with an exception of Li (2001b). Our approach is based on well established theoretical foundation of data compression. Thus, we avoid intensive computation by selecting properly parameters up front. Our algorithm runs in O(N) steps where N is the length of the sequence.

Let us briefly summarise our methods and findings. We partition a DNA sequence (e.g., *chromosome 9*) into fixed length blocks. Guided by universal data compression, we follow Shamir and Costello (2000) and Shamir and Merhav (1999) and set the length of the block to minimise the average redundancy (MDL principle), which the authors of Shamir and Costello (2000) computed to be $(\log N)^{1+\delta}$ ($\delta > 0$) where N is the length of the DNA sequence in base pairs (bps). Then, we invoke the Stein-Ziv lemma (Ziv, 1988) (hypothesis testing and universal data compression) and apply asymptotically optimal discriminant to determine whether two blocks, say b_{i-1} and b_{i+1}, are generated by the same source. It turns out that the optimal discriminant is related to the so-called Jensen-Shannon divergence that was already heuristically used in DNA segmentation (Bernaola-Galván et al., 1996, 1999, 2000). If the discriminant function is positive and large, then a change point (i.e., change of distributions) is expected to occur in block b_i. In this case, we additionally subdivide this block into subblocks of length $(\log \log N)^{1+\delta}$ (i.e., establish level II partition) to assure the best fit of data from the MDL principle point of view. The optimal discriminant function is again applied to these subblocks. Once the change points are found, we compute entropy of these segments (between change points) to identify recurrent sources (see Figure 2 for an illustration).

Figure 2 The optimal discriminant function and entropy for a randomly generated sequence (solid horizontal lines in part (b) are computed from the known distribution)



(a) (b)

This paper is organised as follows. In Section 2, we present our theoretical foundation and describe the proposed segmentation algorithm. In particular, we follow Shamir and Costello (2000) and apply Ziv's analysis to design an asymptotically optimal discriminant that decides whether two blocks are generated by the same source or not. In Section 3, we present our experimental results. We study human *chromosome 9* and *chromosome 20*. We design five different groupings of bases *A*, *C*, *G,* and *T* (e.g., purine vs. pyrimidine, triplets of symbols as in codons) to amplify the discriminant function. This led us to find change points between coding and non-coding regions, starting positions of genes, as well as the beginning of a CpG island.

2 Theoretical underpinning

In this section, we present theoretical underpinning of our approach. Throughout we assume that a sequence $x_1^N = x_1x_2\ldots x_N$ is generated over a finite alphabet A. We sometimes write $x \in A^N$ for x_1^N. As in Shamir and Costello (2000), we partition x_1^N into blocks $b_1, b_2, \ldots, b_K$ such that $|b_i| + \cdots + |b_k| = N$, where $|b_i|$ is the length of the ith block. As shown later we do not lose on asymptotic optimality by making all blocks equal. From now on, we set $|b_i| = b$ for all $1 \le i \le K$ such that $b \cdot K = N$. This partition constitutes *level* I of our algorithm.

The next step is to design an optimal procedure to ascertain that there is a change of distribution in, say block i. We adopt here the approach from the decision theory of optimal hypothesis testing and universal data compression as discussed in Ziv (1988). We first choose an optimisation criterion.

In general, let $x \in A^n$ be a sequence of length n generated either by a source over alphabet A with probability distribution P_1 or a source with probability distribution P_2. At this point, we assume P_1 and P_2 are *known*, and our goal is to decide upon seeing $x \in A^n$, which source generates it so that the probability of 'false decision' (false positives) is exponentially small. More precisely, following Ziv (1988), we are to find a *discriminant* function $\mathrm{d}(x)$, $x \in A^n$ such that

$$P_1(\mathrm{d}(x) > 0) := P_1(x \in A^n : \mathrm{d}(x) > 0)$$

is maximised *subject to*

$$P_2(\mathrm{d}(x) > 0) := P_2(x \in A^n : \mathrm{d}(x) > 0) \le 2^{-\lambda n}$$

for some $\lambda > 0$. Stein's lemma described below generalised by Ziv provides a solution to this problem.

Let us define the *divergence* (rate) $D(P_1||P_2)$ as

$$D(P_1 \,||\, P_2) = \lim_{n\to\infty} \frac{1}{n} \sum_{x\in A^n} P_1(x) \log \frac{P_1(x)}{P_2(x)}$$

if it exists. To assure its existence, we assume hereafter that the source is strongly mixing. In a strongly mixing source the probability of two events A and B defined on two substrings separated by g symbols is bounded as follows:

$$(1 - \psi(g))P(A)P(B) \le P(AB) \le (1 + \psi(g))P(A)P(B)$$

where the mixing coefficient $\psi(g) \to 0$ as $g \to \propto$ (see Szpankowski (2001)). The reader may assume that $x \in A^n$ is generated by a Markov source.

Lemma 1 (Stein, 1952; Ziv, 1988): *Let P_1 and P_2 be generated by a strongly mixing source, and let the discriminant d(x) be such that*

$$P_2(\mathrm{d}(x) > 0)) \leq 2^{-\lambda n}, \quad x \in A^n \tag{1}$$

for some λ>0. Then

(i) For some ε > 0

$$\lim_{n\to\infty} P_1(d(x) > 0) \geq 1 - \varepsilon, \quad x \in A^n \tag{2}$$

if and only if

$$D(P_1 \| P_2) > \lambda \tag{3}$$

(ii) If equation (3) holds, then equation (2) is true for all $\varepsilon > 0$ provided $d(x)$ is optimal and given by

$$d(x) = \frac{1}{n} \log \frac{P_1(x)}{P_2(x)} \lambda \tag{4}$$

for $x \in A^\mathrm{n}$.

The Stein-Ziv lemma gives the best discriminant function for the above optimisation problem provided both distributions P_1 and P_2 are *known*. In reality, the distributions are not known but we can estimate them from data. Our goal now is to find a computable discriminant function that is still asymptotically optimal under the above criteria.

Ziv (1988) constructed such an asymptotically optimal discriminant function. We follow Shamir and Costello (2000) and modify it slightly for our purposes. First, Ziv observed that in order to obtain such a discriminant the training sequence, say $y \in A^n$ must be such that its length is $\Theta(|x|)$ (see Theorem 3 in Ziv (1988)). We simplify our presentation by assuming that $|x| = |y| = n$. Let now $x, y \in A^n$ be two sequences of length n. Define the empirical entropies $H(x)$ $H(x, y)$ as

$$H(x) = -\sum_{a\in A} \frac{n_x(a)}{n} \log \frac{n_x(a)}{n},$$
$$H(x,y) = -\sum_{a\in A} \frac{n_x(a)+n_y(a)}{2n} \log \frac{n_x(a)+n_y(a)}{2n},$$

where $n_x(a)$, $n_y(a)$ are the number of times symbol $a \in A$ occurs in $x, y \in A^n$. Observe that $H(x, y)$ is the empirical entropy of the concatenation $x \cdot y$ of length $2n$. Ziv (1988) in Theorem 4 basically proves that

$$d(x,y) = H(x,y) - \frac{1}{2}H(x) - \frac{1}{2}H(y) - \frac{1}{2}, \quad x, y \in A^n \tag{5}$$

is asymptotically optimal discriminant (in the sense of Lemma 1) as $n \to \infty$.

In view of the above we define a new metric, as in Shamir and Costello (2000),

$$M(x,y) = H(x,y) - \frac{1}{2}H(x) - \frac{1}{2}H(y), \quad x, y \in A^n. \tag{6}$$

If $M(x, y)$ is positive and large, then with high probability, x and y come from different distributions *and* the probability of false positive is exponentially small (with the exponent proportional to the value of $M(x, y)$). Observe also that the bigger $M(x, y)$ is, further apart the distributions are.

Let us come back to our original problem. We recall that the sequence x_1^N is partitioned into blocks b_i of equal size b. In order to decide whether there is a change in distribution in block b_i, we compute the following quantity, as suggested in Shamir and Costello (2000) and Clote and Backofen (2000)

$$M(b_i) = H(b_{i-1}, b_{i+1}) - \frac{1}{2}H(b_{i-1}) - \frac{1}{2}H(b_{i+1}) \tag{7}$$

for $2 \leq i \leq K - 1$. As argued above, if $M(b_i)$ is large, then one expects a change point (change of distribution) to occur in block b_i. We also know that this is the best discriminant in the sense of Lemma 1.

Let us illustrate our approach on an example. In Figure 2, we generated randomly a binary sequence (length $N = 6{,}000$ and $b = 200$) with distribution $P_1 = \{0.4, 0.6\}$ till block b_{10}, then with distribution $P_2 = \{0.1, 0.9\}$ up to block b_{20}, and finally we are again back to the original distribution P_1. Figure 2 clearly indicates change points at b_{10} and b_{20}. We observe that the values of $M(b_{10})$ and $M(b_{20})$ are *not* the same, as it should be since $D(P_1 \| P_2) \neq D(P_2 \| P_1)$. Also, in Figure 2 we plot the entropy in each segment (b_1, b_{10}), (b_{11}, b_{20}), and (b_{21}, b_{30}). Observe that based on the entropy figure, we conclude that we return to distribution P_1 after block b_{20} (the so-called recurrent hidden sources or recurrent segments).

Before we address the issue of selecting the optimal block size b, let us modify the algorithm as suggested in Shamir and Costello (2000). The point is that the block size b may be quite large (we need to have it large in order to take advantage of asymptotic optimality), while we would like to pin down the exact location of the distribution change. Therefore, we introduce *level II* partition. Every block of level I with *high* metric M is further partitioned into blocks β_j of equal length $\beta = |\beta_j|$. For level II, we apply the same metric $M(\beta_j)$ in the same manner as in level I adopted to β partition of the selected blocks of level I. In the next section (see Figures 3–8), we compute metric M for *chromosome 9* and *chromosome 20*.

We still need to decide how to choose the block sizes b of level I and β of level II. This is an important computational problem since selecting all possible block sizes will lead to computational explosion as in Bernaola-Galván et al. (1996). We solve this problem by appealing to the minimum description length (MDL) principle. That is, we choose b and β such that the description of the sequence x_1^N is the smallest possible, or equivalently x_1^N is optimally compressed. Citing Shamir and Costello (Proposition 2 of Shamir and Costello (2000)), we conclude that

$$b = (\log N)^{1+\delta} \tag{8}$$

$$\beta = (\log \log N)^{1+\delta} \tag{9}$$

for some $\delta > 0$. In Figures 5(a) and 6, we compare metric M computed for *chromosome 20 q12-13.12* (*AL021578*) for the *optimal* selection of b and β (see Figure 5(a)) and arbitrary selection of b and β (see Figure 6). Clearly, optimal values of b and β are able to pick up biologically meaningful change points (in this case a CpG island). We also observe that with such selections of the block sizes, the segmentation algorithm runs in $O(N)$ time.

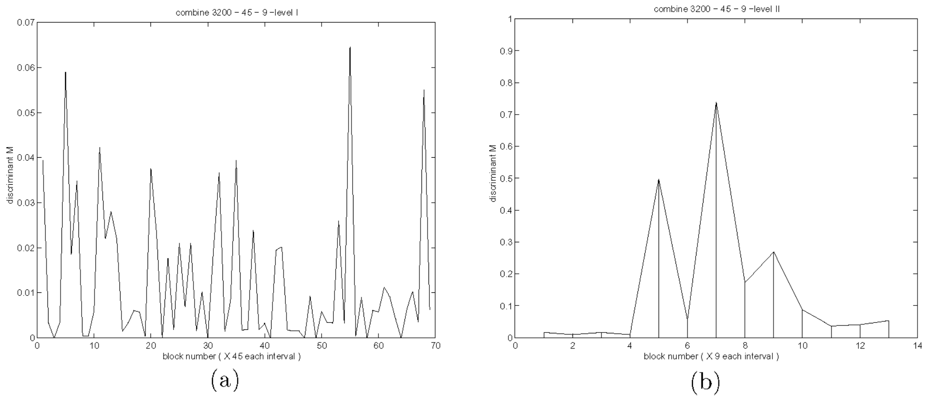
Finally, we comment on the threshold value M_{th} of the discrimination function $M(x, y)$. The idea is to find M_{th} such that for a given α (e.g., $\alpha = 10^{-4}$), the false positive probability does not exceed α, assuring high probability of success (i.e., inequalities (1) and (2) of Lemma 1 hold). From (1), we find λ_{th} such that $2^{-\lambda_{\text{th}} b} = \alpha$ for the block size b. We then set $d(x, y) = M_{\text{th}} - 1/2\,\lambda_{\text{th}} \gg 0$, and by equation (5) we may choose

$$M_{\text{th}} > \frac{1}{2}\lambda_{\text{th}} = \frac{\log_2 \alpha^{-1}}{2b}.$$

For example, for $\alpha = 10^{-4}$ and $b = 300$, we find $M_{th} \geq 0.022$.

In passing, we should point out that metric $M(x, y)$, known also as the Jensen-Shannon divergence, was used before for DNA segmentation in Bernaola-Galván et al. (1996, 1999, 2000). However, the authors of Bernaola-Galván et al. (1996, 1999, 2000) did not use the MDL principle that resulted in much higher algorithmic complexity. Furthermore, in our opinion no proper mathematical justification was presented in Bernaola-Galván et al. (1996, 1999, 2000).

Figure 3 Level I (part a) and level II (part b) of the discriminant function M for the *human chromosome 9q34.2–34.3* from 12,379 bps to 15,578 bps of purine (A and G) and pyrimidine (T and C) with block size 45 bps. The coding region starts around block 54 in level I and around block 7 in level II which corresponds to 2,421 bps – the exact position of the beginning of the coding region



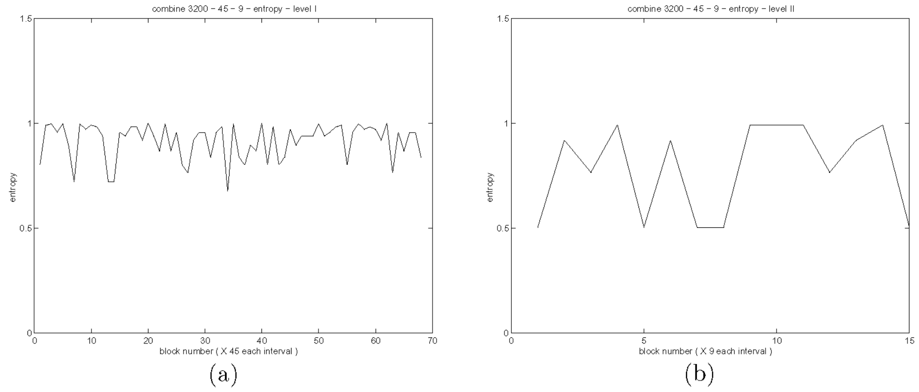
3 Experimental result

In this section, we discuss some experimental studies in which the proposed algorithm was applied to *chromosome 9* and *chromosome 20* in order to identify biologically meaningful change points such as CpG islands, start point of genes, and non-coding regions.

Chromosome 9 is approximately 145 Megabases in length, and chiasma counts suggest a genetic length of 116 cM in males. The Genethon genetic map of *chromosome 9* contains 189 markers and estimates the distance as 138.5 cM in males and 194.5 cM in females, although the distance from the most distant markers to the telomeres is not known. A comprehensive map in SIGMA produced at the *chromosome 9* Workshop in 1996 contains 137 genes (excluding ESTs) and 783 DNA segments. The most recent RH map (Gene-Map 98) shows 1,248 distinct gene based STSs on *chromosome 9*. One of the earliest recognised autosomal linkage groups, that of the ABO blood group locus and the disease Nail-Patella syndrome, is on 9q34, and recent years have seen the identification of the NPS gene as well as several others among the 53 'disease genes' recognised on this chromosome (see http://www.gene.ucl.ac.uk/chr9/).

Firstly, we use clone RP11-100C15 on chromosome 9q34.2–34.3. This portion of DNA contains coding and non-coding regions (i.e., 'junk DNA') as well as CpG islands. We apply our discriminant function to discover these regions. We first observe that working with the four letter alphabet {*A*, *C*, *G*, *T*} is non- desirable due to high variability of *M*. This was already observed by others (e.g., Bernaola-Galván et al., 1996, 1999, 2000), and therefore we either group some letters to form a smaller alphabet or to extend the alphabet.

Figure 4 The entropy of the human chromosome 9q34.2-34.3 from Figure 3 for level I and II

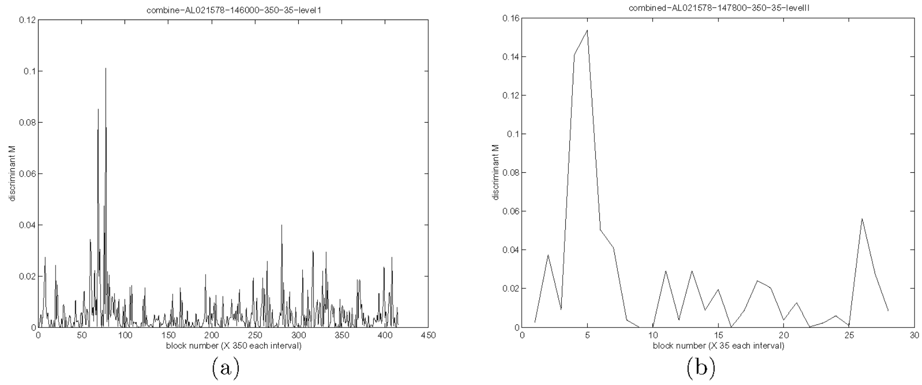


In our first experiment, we consider the portion of chromosome 9q34.2–34.3 from 12,379 bps to 15,578 bps. It contains the noncoding region (i.e., 'junk DNA') from 12,379 to 14,800 bps and the coding region (i.e., exons and introns) of length 778 (14,801–15,578 bps). We apply our discriminant function with block length $b = 45$ in level I and $\beta = 9$ bps in level II for the two symbol alphabet $\{P, R\}$, where $P = \{A, G\}$ (purine) and $R = \{T, C\}$ (pyrimidine). Results are shown in Figure 3. We observe that the highest value of M in level I (see Figure 3(a)) occurs in the 54th block (= 2,409 bps), which approximates very well the change point between coding and noncoding regions at 2,421 bps. In Figure 3(b), we plot the discriminant *M* for level II around the 54th block. The highest value of *M* appears around the 7th block corresponding to 2,426 bps which exactly matches the change point at 2,421.

In order to identify recurrent segments of Figure 3, we plot in Figures 4 the empirical entropy for level I and level II, respectively. The entropy figure is, unfortunately, more fuzzy, especially in level I. This is not a big surprise since, as we computed, $P(R) \approx 0.484963$ in the first segment changes to $P(R) \approx 0.376813$ in the second segment, finally setting at $P(R) \approx 0.483660$. Thus, the difference is rather small and the entropy is close to 1 in all segments. However, the change of entropy is more visible in level II as Figure 4 shows.

In our second set of experiments, we use the human DNA sequence from clone RP3-453C12 of *chromosome 20* in the range q12-13.12. This DNA sequence contains two isoforms of the MATN4 (matrilin 4) gene, the RBPSUHL (recombining binding protein suppressor of hairless-like (Drosophila)) gene, the SDC4 gene, part of a gene for a novel protein (similar to Drosophila CG11753), part of the gene for a p53 response element (clg01), parts of two isoforms of a novel gene for an uncharacterised hypothalamus protein, a gene for a novel protein, ESTs, STSs, GSSs, and CpG islands. Totally it has 1,47,800 bps of DNA sequence data (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&cmd=search&term=AL021578).

Figure 5 Level I (part a) and level II (part b) of the discriminant function *M* for the human *chromosome 20 q12-13.12 (AL021578)* with 1,47,800 bps of purines (A and G) and pyrimidines (T and C) with a CpG island starting at the 80th block of level I and around block 5 of level II. The block size of level I is 350 while level II is 35



(a)

(b)

We first consider 1,47,800 bps of *chromosome 20q12-13.12*. It contains a CpG island starting at position 27,752. In Figure 5, we plot level I and level II of the discriminant function *M* for the alphabet $\{P, R\}$. We clearly identify a transition occurring at the 80th block, which corresponds to the position 27,825 bps. A more careful analysis of level II (Figure 5(b)) reveals that the change point is around 27,725 bps, which is very close to 27,752 bps where the CpG island starts.

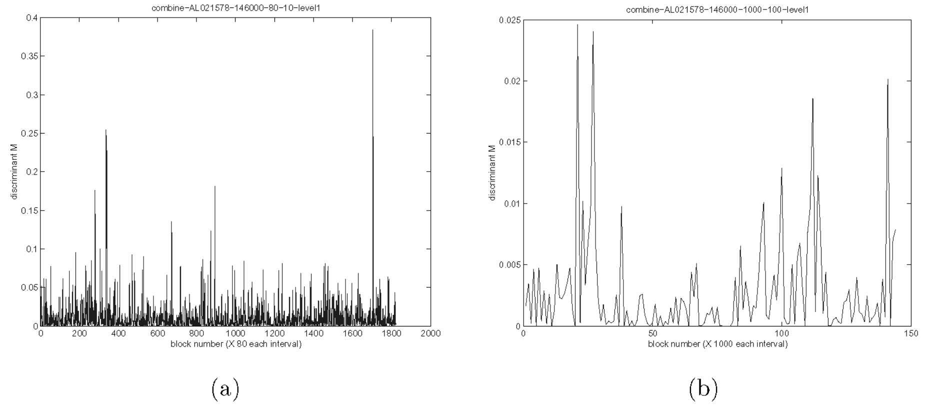
In order to verify the optimality of the block selection ($b = 350$), we repeat experiments of Figure 5 but with a non-optimal block size selection. In Figure 6, we re-plot Figure 5(a) with the block sizes $b = 80$ and $b = 1{,}000$. Clearly, non-optimal block sizes make the graph of M more fuzzy, and one cannot identify biologically meaningful change points (e.g., the CpG island of Figure 5).

In the previous experiments, we grouped $\{A, C, G, T\}$ into a new alphabet of size two $\{P, R\} = \{A/G, C/T\}$ (i.e., purines and pyrimidines). We refer to this approach as the *2-phase method*. There is, however, another way of grouping $\{A, C, G, T\}$ that is particularly useful for finding CpG islands. In Figure 7, we group C and G together, and then A and T together, for the AL021578 sequence. There are CpG islands in the intervals (4,751, 5,279), (9,429, 9,829), (22,943, 23,233), (27,905, 28,288), (59,288, 59,896), and (1,27,233, 1,27,761). In Figure 7 we identify six out of 16 CpG islands in AL021578.

We now turn our attention to discovering coding regions in *chromosome 20*. When studying such translated regions, it makes sense to group symbols in triples (since, as we know, they are responsible for coding proteins). It was observed (Bernaola-Galván, 2000) that such triples do have distinctly new distribution in the coding region. Therefore, we propose three additional methods in which we expand the alphabet rather than shrink it:

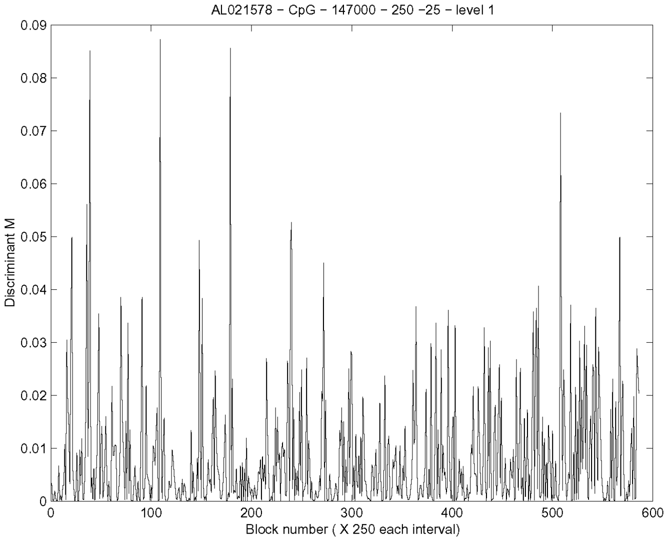
- *12-phase method.* In this case, we consider the original alphabet $\{A, C, G, T\}$ but to each symbol we assign one of three phases $j = 0, 1, 2$, that is, a nucleotide at position i has its phase $j = i \bmod 3$. Therefore, every nucleotide of the DNA sequence can be substituted by one of the following symbols: A0 A1 A2 T0 T1 T2 G0 G1 G2 C0 C1 C2. For instance, A2 means that we have found a nucleotide A with phase equal to 2. This method was already used in Bernaola-Galván (2000).
- *24-phase method*. In this method, we consider triplets of purines and pyrimidines (hence $2^3 = 8$ basic cases) together with three phases as in the *12-phase method*. This gives an extended alphabet of size 24.
- *192-phase method*. The same as the *24-phase method* except that we use the original alphabet $\{A, C, G, T\}$ (i.e., $192 = 3 \cdot 4^3$).

Figure 6 Level I of the discriminant function M (with block size 80 bps and 1,000 bps, respectively) for the human *chromosome 20q12-13.12(AL021578)* of purines (A and G) and pyrimidines (T and C). The CpG island starting at 27,752 in Figure 5(a) cannot be identified here



(a) (b)

Figure 7 Six CpG islands found in DNA sequence AL021578 using our method with the block size equal to 250. The change points are at positions 21, 39, 91, 112, 239, and 510, each being the beginning of a CpG island



In addition, the above three methods are enhanced with a 'shifting value' representing the beginning of the first block in level I. That is, the shifting value specifies the starting position of the DNA reading frame. For instance, if the shifting value is 300 and the block size $b = 600$, then the first block is from 300 to 900. Obviously, different shifting values result in different data in each block leading further to new values of M. This allows us to identify 'unexpected' change points.

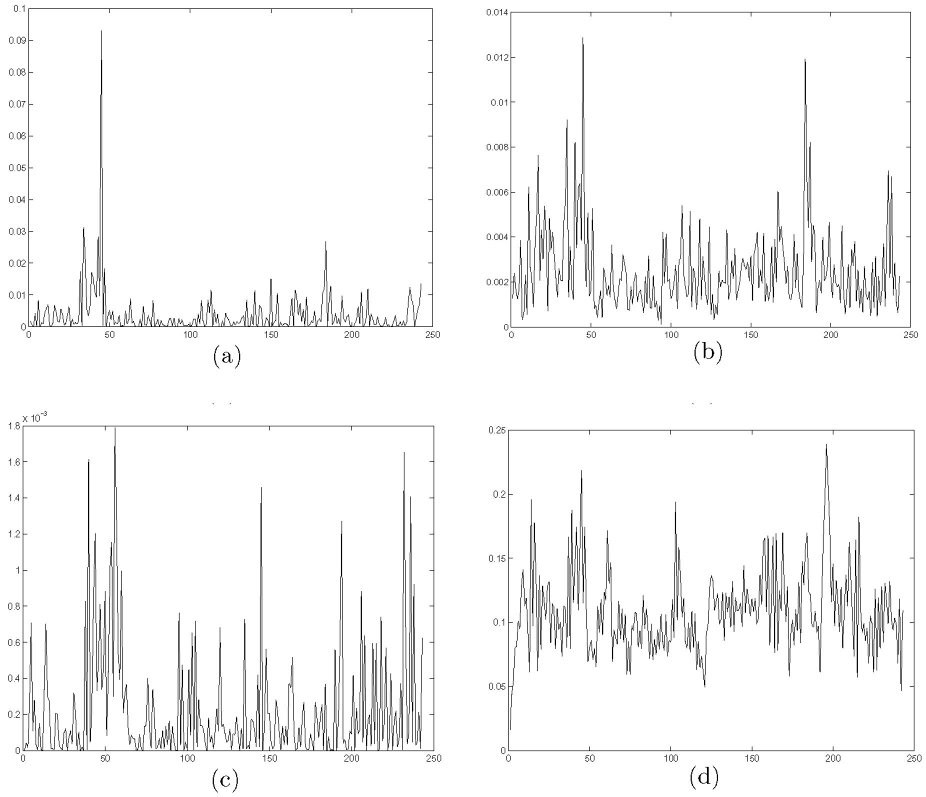
Table 1 Top two transition points in *chromosome 20q12-13.12* (AL021578)

Shifting num	*2 phase*	*12 phase*	*24 phase*	*192 phase*
0	45/39	60/52	45/38	45/150
100	180/45	60/48	45/38	194/45
200	45/180	236/60	180/39	194/45
300	45/180	60/51	45/180	14/45
400	45/38	60/236	45/180	45/218
500	45/38	60/51	45/180	45/195
600	45/38	60/39	45/36	195/45

In Table 1, we list two highest transition points for the four methods we just discussed together with various shifting values in the range 0–600. As before, we analyse the DNA sequence of human chromosome *20q12-13.12*. In Table 1, we show the block numbers for each experiment (e.g., 45/180 means that the first transition point is at the 45th block and the second highest transition point is at the 180th block). Finally, we select biologically meaningful transition points. The results are presented in Figure 8 from which we conclude:

- In the *2-phase method* with shifting value 400, the second highest transition point is at 38th block that corresponds to the position 22,901 bps. There is a gene (RBPSUHL) that starts at position 22,937 in the DNA sequence.
- In the *24-phase method* with shifting value 300, the transition point is at 45th block (i.e., position 27,001 bps). There is a CpG island beginning at position 27,050.
- In the *12-phase method* with shifting value 300, the highest transition point is at the 61st block corresponding to 36,601 bps. There is a gene (SDC4) that starts at position 36,558 in the DNA sequence.
- In the *192-phase method* with shifting value 200, the transition point is at the 195th block, that is, at position 1,27,001 bps. According to the DNA database, there is a gene (dj453C12.4) that begins at position 1,27,089.

Figure 8 The discriminant function *M* for the human chromosome *20 q12-13.12* (a) *2-phase method* with the shifting value 400; a gene (RBPSUHL) starts at the position close to 38th block; (b) *24-phase method* with shifting value 300; a CpG island begins near the 45th block; (c) *12-phase method* with shifting value 300; gene (SDC4) starts at the position near 61st block and (d) *192-phase method* with shifting value 200; gene (dj453C12.4) begins at the position near the 195th block



4 Conclusion

In conclusion, we use the algorithm and analysis of Shamir and Costello (2000) to present rigorous justifications of applying the discrimination function $M(x, y)$ for DNA segmentation. We base our analysis on the Stein-Ziv lemma, and the MDL principle that guide us in the selection of block sizes. We also use several different grouping methods and shifting values to magnify change points. Experimental results confirm our theoretical predictions and coincide with biologically meaningful transition points (e.g., coding/non-coding regions, CpG islands, and starting points of genes).

Our coding/noncoding region findings suggest that:

- The *2-phase method* and the *24-phase method* produce similar results, and shifting value does not alter significantly results (e.g., the top points are around the 38th, the 45th, and the 180th block).
- Shifting helps in the *12-phase method* and the *192-phase method* (Table 1). In particular, combining A/G and C/T leads to more stable results (e.g., the transitions in the 61st and the 195th block were found, by the *12-phase method* and the *192-phase method,* respectively).
- Sometimes the first highest calculated value has biological meaning, while in other cases the second highest value leads to a biologically meaningful result. By trying various methods and shifting values, we are usually able to find interesting change points.
- For finding coding regions, we conclude that grouping the original alphabet in triplets, as in the *24-phase method* and *192-phase method*, is better than the *12-base method* proposed in Bernaola-Galván (2000).

In summary, our observations suggest that the proposed segmentation algorithm is able to identify biologically meaningful regions. Its running time is $O(N)$, which is very attractive for biological applications.

Acknowledgement

We thank Dr. G. Shamir for useful discussions about piecewise stationary sources. Preliminary results were presented at the Computational Systems Bioinformatics, Stanford in 2003. This author's research was supported in part by NSF Grant CCR-0208709 and the NIH Grant R01 GM068959-01.

References

Baldi, P. and Brunak, S. (2000) *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, Mass.

Bernaola-Galván, P., Carpena, P., Roman-Roldán, R. and Oliver, J.L. (1999) 'Decomposition of DNA sequence complexity', *Phys. Rev. Letters*, Vol. 83, pp.3336–3339.

Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J.L., Roman-Roldán, R. and Stanley, H.E. (2000) 'Finding borders between coding and noncoding DNA', *Phys. Rev. Letters*, Vol. 85, pp.1342–1345.

Bernaola-Galván, P., Roman-Roldán, R. and Oliver, J.L. (1996) 'Compositional segmentation and long-range fractal correlations in DNA sequences', *Phys. Review E*, Vol. 53, pp.5181–5189.

Churchill, G.A. (1989) 'Stochastic models for heterogeneous DNA sequences', *Bull. Mathematical Biology*, Vol. 51, pp.79–94.

Clote, P. and Backofen, R. (2000) *Computational Molecular Biology*, John Wiley & Sons, Chichester.

Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, John Wiley & Sons, New York.

Crowley, E., Roeder, K. and Bina, M. (1997) 'A statistical model for locating regulatory regions in genomic DNA', *J. Mol. Biol.*, Vol. 268, pp.8–14.

http://www.gene.ucl.ac.uk/chr9/.

Li, W. (2001a) 'DNA segmentation as a model selection process', *Proc. RECOMB*, pp.204–210.

Li, W. (2001b) 'New stopping criteria for segmenting DNA sequences', *Phys. Rev. Letters*, Vol. 86, pp.5815–5818.

Nevill-Manning, C. and Witten, I. (1999) 'Protein is incompressible', *Proc. Data Compression Conference*, pp.257–286.

Pevzner, P. (2000) *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, Cambridge.

Ramensky, V.E., Markeev, V., Roytberg, M. and Tumanyan, V. (2000) 'DNA segmentation through the Bayesian approach', *J. Computational Biology*, Vol. 7, pp.215–231.

Sahinalp, S., Eichler, E., Goldberg, P., Berebrink, P., Friedetzky, T. and Egrin, F. (2002) *Statistical Identification of Uniformly Mutated Segments within Repeats*, CPM 2002, LNCS 2373, pp.249–261.

Salzberg, S., Searls, D. and Kasif, S. (1999) *Computational Methods in Molecular Biology*, Elsevier, Amsterdam.

Shamir, G. and Costello, D. (2000) 'Asymptotically optimal low-complexity sequential loss-less coding for piecewise-stationary memoryless sources. I: the regular case', *IEEE Trans. Information Theory*, Vol. 46, pp.2244–2467.

Shamir, G. and Merhav, N. (1999) 'Low-complexity sequential lossless coding for piecewise-stationary memoryless sources', *IEEE Trans. Information Theory*, Vol. 45, pp.1498–1519.

Szpankowski, W. (2001) *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York.

Tabus, I., Korodi, G. and Rissanen, J. (2003) 'DNA sequence compression using the normalized maximum likelihood model for discrete regression', *Proc. Data Compression Conference*, Snowbird, pp.253–262.

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&cmd=search&term=AL021578.

Ziv, J. (1988) 'On classification with empirically observed statistics and universal data compression', *IEEE Trans. Information Theory*, Vol. 34, pp.278–286.