An Optimal DNA Segmentation Based on the MDL Principle *

April 1, 2003

Wojciech Szpankowski and Wenhui Ren
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
{spa,renwh}@cs.purdue.edu

Abstract

A major challenge facing computational biology is the post-sequencing analysis of genomic DNA and other biological sequences. Biological world is highly stochastic in its behavior and inhomogeneous. Indeed, 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. The transition between homogeneous and inhomogeneous regions, known also as change points, carry important biological information. Computational methods used to identify these homogeneous regions are called segmentation. We will view a DNA sequence as a non-stationary processes and apply recent novel techniques of universal coding to discover stationary (possibly recurrent) segments. Our goal is to employ rigorous methods of information theory to quantify structural properties of DNA sequences. In particular, we adopt the Stein-Ziv lemma to find asymptotically optimal discriminant function that determines whether two DNA segments are generated by the same source and assuring exponentially small false positives. Then we apply the Minimum Description Length (MDL) principle to select parameters of our segmentation algorithm so that the underlying DNA sequence has the smallest possible length (best compression). The MDL principle is adopted since recent analyses indicate that biological sequences are well compressed (by nature). Finally, we perform extensive experimental work on human chromosome 22. In particular, we observe that grouping A and G (purines) and T and C (pyrimidines) leads to better segmentation of chromosome 22 and identification of biologically meaningful transition points.

Key Words: DNA segmentation, stochastic modeling, universal data compression, MDL principle, model selection, piece-stationary sequences.

One good theory is worth a thousand computer runs.
Modern version of Confucius (D. G. Luenberger)

*This research was supported in part by NSF Grants CCR-9804760 and CCR-0208709.
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. For example, it is known that a typical DNA sequence is not homogeneous and that (approximate) repetitions as well as regions with varying statistical properties can have biological meanings (e.g., regulatory elements, structural features of the DNA, CpG islands, coding versus non-coding regions [14, 18]). On the other hand, information theory [7, 20], in particular source coding better known as data compression, explored over the last 50 years 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 modeling part that finds a fit between data and a statistical model (e.g., based on the Minimum Description Length or in short the MDL principle). This is also known as probability assignment and it is common to many pattern discovery algorithms. We apply rigorous methods of information theory to DNA segmentation and finding recurrent sources in a DNA sequence (cf. also [4, 10, 11, 17]).

Probabilistic methods such as Hidden Markov models, Bayesian approaches, large deviations, and combinatorial tools based on graphs, enumeration methods and combinatorial optimizations are recognized as useful tools in computational biology [1, 6, 14]. Needless to say, the biological world is highly stochastic in its behavior and inhomogeneous. We adopt a novel information theoretical approach to analyze DNA sequences which are assumed to be generated by a non-stationary source. As shown in Figure 1 (and discussed below) molecular sequences exhibit different statistical properties in various parts of DNA (e.g., coding region versus non-coding). Following recent experimental results [13, 21], 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.

To justify our approach from biological point of view, we briefly review some basic facts that guide us and serve us for our future research. In particular, we present examples of non-stationary nature of biological sequences. We refer to Figure 1 (cf. [18]).

- Oligomer frequency can be used to distinguish between coding and noncoding regions. We observe that coding regions are rather short (averaging 150 nucleotides in length). Staden and McLachlan observed 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. We aim at applying tools of piece-stationary sequences [12, 19] and the Stein-Ziv lemma [22] to find these boundaries (cf. [4]).

- Human gene mutations have been observed more frequently at certain dinucleotides than others. For example, in the dinucleotides 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. Hidden Markov models are currently used to find such and other islands (cf. [8, 14]). We shall apply the piece-stationary model discussed above and some recent precise results of analysis of algorithms (cf. [16, 20]) to find these islands.

- 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. This necessitates exact or very precise asymptotics of the pattern matching occurrences to overcome false positives. Again, non-stationary approach proposed here together with the methods of [16] can indicate a transition between these regions.

- 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*, known also 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 Sueoka 1959 (base composition), Elton 1974 (bacteria), Rouchka et. al. 1997 (CpG islands), Churchill 1989 [5] (hidden Markov model), Fickett 1992 (walking Markov model), and others (e.g., moving window approach). Techniques used range from Bayesian approach [15, 17], model selection [10, 11], and entropy-related techniques [2, 3, 4]. One particular attractive segmentation is the divide-and-conquer approach [2] in which DNA is first segmented into two substrings for which the base composition is maximized, and this is carry on the left and right substrings. The technique is computationally involved and there is no
clear criteria for stopping the recursion. The latter problem is addressed in [11]. Our approach is based on well established theoretical foundation of data compression so we avoid intensive computation by selecting properly parameters, as discussed below.

Let us briefly summarize our methods and findings. We partition a DNA sequence (e.g., *chromosome 22*) into fixed length blocks. Guided by universal data compression, we shall follow Shamir and Costello [19] and set the length of the block to minimize the average redundancy (MDL principle) which turns out 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 [22] (of hypothesis testing and universal data compression) and apply asymptotically optimal discriminant to determine whether two blocks, say \(b_i\) 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 (cf. [2, 3, 4]). If the discriminant function is positive, then a change point (i.e., change of distributions) between blocks \(b_i\) and \(b_{i+1}\) is expected. Since the length of \(b_i\) is of order \(O(\log N)\) bps, we then subdivide the block into subblocks of length \((\log \log N)^{1+\delta}\) (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 (hidden) sources that generate (segments of) the sequence (cf. Figure 2 for an illustration).

This paper is organized as follows. In the next section we present our theoretical foundation and describe a segmentation algorithm. In particular, we follow Shamir and Costello [19] and slightly modify Ziv’s analysis to design an asymptotically optimal discriminant to decide whether two blocks are coming from the same source or not. In Section 3 we present our experimental results. We study the human *chromosome 22* which is the second smallest of the human autosomes (e.g., it is known that the short arm (22p) of this chromosome contains a series of tandem repeat structures including the array of genes that encode the structural RNAs of the ribosomes). We take some 160000 bps of chromosome 22 and divide it into blocks of size 400 bps. Applying our discriminant we clearly identify segments as well as hidden sources that generate them. In particular, we observe that grouping A and G (purines) and T and C (pyrimidines) leads to a better segmentation of chromosome 22 and identification of biologically meaningful transition points.

Finally, we should point out that in this conference version of the paper we rather concentrate on presenting rigorous methods of analysis than showing their experimental relevance to finding particular segments in a DNA (being it coding/noncoding regions or CpG islands). This puts squarely our contribution between *quantitative biology* and computational biology (efficient segmentation algorithms).

## 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 \(\mathcal{A}\). We sometimes write \(x \in \mathcal{A}^N\) for \(x_1^N\). As in [19], we partition \(x_1^N\) of length \(N\) (e.g., \(N\) bps) into blocks \(b_1, b_2, \ldots, b_K\) such that \(|b_1| + \cdots + |b_K| = N\), where \(|b_i|\) is the length of the \(i\)th block. As shown below we are
not losing on asymptotic optimality by making all blocks equal. From now on we set \( |b_i| = b \)
for all \( 1 \leq i \leq K \) 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 [22]. We first choose an optimization criterion. In general, let \( x^n \) be a sequence generated either by a source 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 \( x^n \) which source generates it so that the probability of “false decision” (false positives) is exponentially small. More precisely, we want to find a discriminant function \( d(x) \), \( x \in \mathcal{A}^n \) such that
\[
P_1(d(x) > 0) := P_1(x \in \mathcal{A}^n : d(x) > 0)
\]
is maximized subject to
\[
P_2(d(x) > 0) := P_2(x \in \mathcal{A}^n : d(x) > 0) \leq 2^{-\lambda n}
\]
for some \( \lambda > 0 \). Stein’s lemma generalized by Ziv provides a solution to this problem.

Let us define the divergence (rate) \( D(P_1||P_2) \) as
\[
D(P_1||P_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{x \in \mathcal{A}^n} P_1(x) \log \frac{P_1(x)}{P_2(x)}
\]
if it exists. To assure its existence we assume from now one that the source is strongly mixing (cf. [20] for a definition). The reader may assume that \( x^n \) 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 be such that
\[
P_2(d(x) > 0)) \leq 2^{-\lambda n}
\]
for some \( \lambda > 0 \).

(i) For some \( \varepsilon > 0 \)
\[
\lim_{n \to \infty} P_1(d(x) > 0) \geq 1 - \varepsilon
\]
if and only if
\[
D(P_1||P_2) > \lambda.
\]
(ii) If (3) holds, then (2) is true for all \( \varepsilon > 0 \) provided \( d(x) \) is optimal and given by
\[
d(x) = \frac{1}{n} \log \frac{P_1(x)}{P_2(x)} - \lambda
\]
for \( x \in \mathcal{A}^n \).
The Stein-Ziv lemma gives the best discriminant function for the above optimization problem provided both distributions \( P_1 \) and \( P_2 \) are known. In real world 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 in [22] constructed such an asymptotically optimal discriminant function. We shall slightly modify it for our purposes, as already suggested in [19]. First, Ziv observed that in order to obtain such a discriminant the training sequence, say \( y \) must be such that its length is \( O(|x|) \) (cf. Theorem 3 in [22]). We simplify our presentation by assuming that \(|x| = |y| = n\). Let now \( x, y \in \mathcal{A}^n \) be two sequences of length \( n \). Define the empirical entropies \( H(x) \) \( H(x, y) \) as

\[
H(x) = - \sum_{a \in \mathcal{A}} \frac{n_x(a)}{n} \log \frac{n_x(a)}{n},
\]

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

where \( n_x(a) \) is the number of times symbol \( a \in \mathcal{A} \) occurs in \( x \in \mathcal{A}^n \). Observe that \( H(x, y) \) is the empirical entropy of the concatenation \( x \cdot y \) of length \( 2n \). Ziv [22] in Theorem 4 basically proves that

\[
d(x, y) = H(x, y) - \frac{1}{2} H(x) - \frac{1}{2} H(y) - \frac{1}{2} \lambda,
\]

is asymptotically optimal discriminant as \( n \rightarrow \infty \).

In view of the above we define a new metric, as in [19],

\[
M(x, y) = H(x, y) - \frac{1}{2} H(x) - \frac{1}{2} H(y), \quad x, y \in \mathcal{A}^n.
\]
If $M(x, y)$ is positive and large, then with high probability $x$ and $y$ are coming 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 [19],

$$M(b_i) = H(b_{i-1}, b_{i+1}) - \frac{1}{2}H(b_{i-1}) - \frac{1}{2}H(b_{i+1}), \quad 2 \leq i \leq K - 1. \quad (7)$$

As argued above, if $M(b_i)$ is large, then one expects a change point (change of distribution) to occur in $b_i$. We also know that this is the best discriminant in the sense discussed above.

Let us illustrate our approach on an example. In Figure 2 we generated randomly a binary sequence (of length 6000 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 went back to the original distribution $P_1$. Figure 2(a) 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(b) we plot the entropy in each segments $(b_1, b_{10})$, $(b_{11}, b_{20})$, and $(b_{21}, b_{30})$. Observe that based on the entropy figure we conclude that we return to distribution $P_1$ after block $b_{20}$ (so called recurrent hidden sources or recurrent segments).

Before we address the issue of selecting optimal block size $b$, let us modify the algorithm, as suggested in [19]. 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) and 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 $\beta_j$ of equal length $\beta = |\beta_j|$. For level II we apply the same metric $M(\beta_j)$ adopted to $\beta$ partition. In the next section (cf. Figures 4–6) we compute metric $M$ for level I and level II for chromosome 22. We shall discuss these results soon.

There is only one question left. How to choose the block sizes $b$ of level I and $\beta$ of level II. We solve this problem by appealing to the Minimum Description Length (MDL) principle. That is, we choose $b$ and $\beta$ such that the description of the sequence $x_1^N$ is the smallest possible. Citing Shamir and Costello (cf. Proposition 2 of [19]) we conclude that

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

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

for some $\delta > 0$. In figures below we compare the metric $M$ for the optimal selection of $b$ and $\beta$ (cf. Figures 4) and arbitrary selection of $b$ and $\beta$ (cf. Figure 6). We also observe that with such selections of the block sizes, the segmentation algorithm runs in $O(N)$ time.

3 Experimental Results and Conclusion

We now discuss our experimental studies where we apply just designed metric and algorithm to chromosome 22. The sequence of human chromosome 22 was completed in December 1999.
Figure 3: The discriminant function $M$ for the human chromosome 22 over all nucleotides $\{A, C, G, T\}$ with 160000 bps and block size of level I equal to 400 bps while block size of level II equal to 50.

An international consortium of sequencing centers released into the public domain the genetic code of the 33.5 million bps that comprise the euchromatic portion of human chromosome 22. This was the first time a human chromosome has been sequenced and included the largest continuous sequence determined from any organism at the time (23 million bps).

Let us first say a few words about chromosome 22. Chromosome 22 is the second smallest of the human autosomes. The short arm (22p) contains a series of tandem repeat structures including the array of genes that encode the structural RNAs of the ribosomes, and is highly similar to the short arms of chromosomes 13, 14, 15 and 21. The long arm (22q) is the portion of human chromosome 22 that contains the protein coding genes and this is the region that has now been sequenced. The completed sequence consisted of 12 contiguous segments covering 33.4 million bps separated by 11 gaps of known size. One of these gaps has subsequently been close by the Oklahoma group.

In all our experiments we take a small portion of chromosome 22q. We do experiments with about 160000 bps that are divided into 400 blocks of length $b = 400$ (level I) and $\beta = 50$ of level II. In Figure 3 we plot the metrics $M(b)$ (cf. Figure 3(a)) and $M(\beta)$ (cf. Figure 3(b)) for four letter alphabet $\{A, C, G, T\}$. One identifies several segments of high value of the metric $M$, however, variability of $M$ is too high to draw any solid conclusion. This was already observed by others (e.g., [2, 3, 4]), and therefore we either group some letters to form a smaller alphabet or extend the alphabet.

Let us group $A$ and $G$ into purine $R = \{A, G\}$ and $T$ and $C$ into pyrimidines $Y = \{C, T\}$, (In the rest of the paper we work with the alphabet $\{R, Y\}$.) In Figure 4 we plot $M(b)$ of level I (see part (a) of the figure) and $M(\beta)$ of level II (part (b) of the figure). After grouping we see (at least) three clear segments in Figure 4(a) that are further analyzed on level II (cf. Figure 4(b)). In particular, $M(b)$ increases rapidly around block 225 in part (a) of the figure.
Figure 4: The discriminant function $M$ for the human chromosome 22 of purines (A and G) and pyrimidines (T and C) over 160000 bps with block size of level I equal to 400 bps and block size of level II equal to 50.

Part (b) suggests that the change point occurs around block $227 = 10800$ bps. From gene map of chromosome 22 at http://www.sanger.ac.uk/cgi-bin/humchr?chr=22 we observe that the first transition point is around 10900 bps, which coincides with our finding. To identify recurrent segments we plot in Figure 5 the empirical entropy (of each block) for level I (part (a) of the figure) and level II (part (b) of the figure). The entropy figure is, unfortunately, more fuzzy, especially on level I. This is not a big surprise since, as we computed, $P(R) \approx 0.484963$ in the first segment, change to $P(R) \approx 0.376813$, finally setting to $P(R) \approx 0.483660$. As one observes the changes are 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 5(b) shows.

Finally, we turn our attention to Figure 6, where we test our selection of parameters $b$ and $\beta$. In general, we argue that too small value of $b$ will lead to imprecise results due to transient behavior, that is, asymptotics do not yet work for too small $b$. If we select too large $b$, potential high values of $M$ are smoothed out (e.g., think of one high value among many small values). In Figure 6 we set $b = 800$ and this seems to be too large since some interesting phenomena of Figure 4 either disappear or are significantly smoothed out.

Concluding, in this paper we present rigorous justifications of using our metric $M$ to DNA segmentation. We base our analysis on the Stein-Ziv lemma, and the MDL principle that guide us in the selection of block sizes. Experimental results seem to confirm our theoretical prediction and they coincide with biologically meaningful transition points. In passing we should point out that the metric $M$, known also as the Jensen-Shannon divergence, was used before for DNA segmentation [2, 3, 4]. However, in our opinion no proper mathematical justification was presented.

\footnote{We should say that we are happy that metric $M$ captures even such some changes in the distribution.}
Figure 5: The entropy of the human chromosome 22 from Figure 4 for (a) level I and (b) level II.

Figure 6: The same situation as in Figure 4 except that the block size of level I is 800 instead of 400 (level II block size is the same as before, i.e., 50).
References


