

# Identification of Consensus Patterns in Unaligned DNA and Protein Sequences: a Large-Deviation Statistical Basis for Penalizing Gaps

key words: large deviations, information theory, sequence alignment, DNA, RNA, protein

Gerald Z. Hertz\* and Gary D. Stormo  
(hertz@colorado.edu; stormo@colorado.edu)

Department of Molecular, Cellular, and Developmental Biology  
University of Colorado  
Boulder, CO 80309-0347 U.S.A.  
(303) 492-1474

running head: large-deviation statistics and multiple sequence alignments

\* To whom reprint requests should be sent

*Proceedings of the 3rd International Conference on Bioinformatics and Genome Research,*  
(H. A. Lim and C. R. Cantor, eds)  
World Scientific Publishing Co., Ltd., Singapore, 1995, pp. 201–216.

# Abstract

Using log-likelihood statistics to compare sequence alignments, we have been able to determine alignments from multiple, unaligned, functionally related, DNA (Stormo and Hartzell. 1989. *Proc. Natl. Acad. Sci. USA* **86**, 1183–1187; Hertz *et al.* 1990. *Comput. Appl. Biosci.* **6**, 81–92) and protein sequences. In this paper, we reanalyze DNA sequences that bind the *E. coli* repressor LexA to demonstrate the ability of our scoring scheme to identify patterns when each sequence can contain zero or more binding sites.

The scoring formula we have used previously does not allow for insertions and deletions in the alignments. In this paper, we use large-deviation statistics to extend the scoring formula to allow for insertions and deletions. The insertion-deletion penalty of this scoring scheme depends exclusively on the observed alignment rather than on previous observations or the user’s intuition. We also describe the close relationship between our formulas and hidden markov models. Finally, we present results of applying this new scoring formula to align a set of *E. coli* promoter DNA sequences.

## 1 Introduction

Molecular biologists frequently can obtain interesting insight by aligning a set of related DNA, RNA, or protein sequences. Such alignments can be used to determine either evolutionary or functional relationships. However, unless the sequences are very similar, it is necessary to have a specific strategy for measuring—or scoring—the relatedness of the aligned sequences. If the alignment is not known, one can be determined by finding an alignment that optimizes the scoring scheme.

Various strategies have been used for scoring an alignment of multiple sequences. For example, the score can be the sum of each of the pairwise scores [1, 14]; the score can be based on the evolutionary path required for the sequences to evolve from a common ancestor [16]; or the score can be based on log-likelihood statistics [10, 20]. Each approach makes different assumptions that are relevant to which approach might be most appropriate. If the sequences are assumed to be conserved because they have not had time to completely diverge since splitting from a common ancestor, then a scoring scheme based on an evolutionary tree is appropriate.

On the other hand, our interest has been in sequence conservation due to functional constraints. Thus, our approach has been to compare different alignments according to their log-likelihood ratio, which ranks alignments according to their probability of occurring by chance without regard to how the alignments might have evolved. This log-likelihood statistic is also related to large-deviation statistics [5], information theory [7], and the thermodynamics of protein binding to DNA [4, 22].

Most scoring schemes for analyzing biological alignments are somewhat arbitrary. Log-likelihood statistics offers an objective method for scoring multiple alignments that is based solely on the observed alignment. However, the log-likelihood formula used in our previous work on determining the alignment of multiple, unaligned, functionally related, DNA [10, 21] and protein (unpublished results) sequences did not allow for insertions and deletions in the alignment. In this paper, we use large-deviation statistics to generalize our scoring formula

	A	A	T	T	G	A
	A	G	G	T	C	C
	A	G	G	A	T	G
	A	G	G	C	G	T
	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

Figure 1: An example of a matrix summarizing a DNA sequence alignment not containing gaps. On the top is an alignment of four 6-mers. Below is a matrix containing the number of times— $n_{i,j}$ —that the indicated letter is observed at the indicated position of this alignment.

to overcome this limitation.

In this paper, we discuss three progressively more general scoring formulas, such that each version is a special case of the succeeding version. The first version is the one we have used previously to compare alignments not containing insertions and deletions [10, 21]. The second version includes the extension that allows for insertions and deletions. And the third version incorporates adjacent correlations. These three scoring formulas correspond to progressively more complicated models for describing a sequence alignment. In a similar fashion, other scoring formulas can be derived from a multitude of other models for describing alignments. Finally, we present results of using these formulas to align DNA sequences that bind the *E. coli* repressor LexA and to align *E. coli* promoter DNA sequences.

## 2 Large-deviation derivations of the Scoring Formulas

### 2.1 The Information Content for a Sequence Alignment NOT Containing Gaps

The first step for determining a scoring formula for an aligned set of sequences is to develop a model for describing the alignment. In this section we describe a simple model which does not contain gaps and in which each position of the alignment is considered to be independent.

Given an aligned set of  $L$ -mers (i.e., sequences of length  $L$ ), summarize the alignment in an  $A \times L$  matrix (e.g., Figure 1). The  $A$  corresponds to the size of the alphabet of interest—e.g., 4 bases in the case of DNA and 20 amino acids in the case of protein: each of the  $A$  rows corresponds to one of the letters of the alphabet. The  $L$  corresponds to the width of the alignment: each of the  $L$  columns corresponds to one of the positions within the alignment. The elements of the matrix are,  $n_{i,j}$ , the number of times that letter  $i$  is observed at position  $j$ . Alternatively, the matrix could contain the frequencies,  $f_{i,j} = n_{i,j}/N$ , where  $N$  is the total number of sequences being summarized in the matrix (i.e.,  $N = \sum_{i=1}^A n_{i,j}$ ).

Each letter  $i$  is assumed to occur with some *a priori* probability  $p_i$  such that  $\sum_{i=1}^A p_i = 1$ . These *a priori* probabilities might be determined from the observed frequencies in the particular data set being analyzed or from some other model. For example, the *a priori* probabilities in the case of a DNA alignment might be the genomic frequencies of the nucleotide

bases. Since each position of the alignment is considered independent, the overall probability of the alignment matrix is the product of the multinomial probability for each column. Therefore, the probability  $P_{\text{matrix}}$  of the alignment matrix is

$$P_{\text{matrix}} = \prod_{j=1}^L \left[ \frac{N!}{\prod_{i=1}^A n_{i,j}!} \prod_{i=1}^A p_i^{n_{i,j}} \right]. \quad (1)$$

We refer to the following formula as the information content of an alignment matrix:

$$I_{\text{matrix}} = \sum_{j=1}^L \sum_{i=1}^A f_{i,j} \ln \frac{f_{i,j}}{p_i}. \quad (2)$$

Equation **2** is a measure of the distance from the center of the distribution where  $f_{i,j} = p_i$ . When  $f_{i,j} = p_i$ , the distance is at a minimum and equals zero. The distance is maximized when the least expected letter occurs exclusively—i.e.,  $f_m = 1$  and  $p_m \leq p_i$  for all values of  $i$ . This formula has gone by various names according to the perspective of those who have derived it. When multiplied by  $-N$ , this formula is the log-likelihood ratio for the multinomial distribution in equation **1**. When motivated by information theory, this formula is called the Kullback-Leibler information [12] or relative entropy [7]. And, when derived from large-deviation principles, it is the *large-deviation rate function* for the distribution in equation **1** [5].

Our ultimate goal is to find the alignment with the greatest statistical significance, where the statistical significance is the inverse of the probability of observing an alignment having the observed information content or greater and having the observed width or less. When the information content is small and the number of sequences is large,  $2NI$  tends to a chi-squared distribution since  $-NI$  is a log-likelihood ratio. In the alignments described in this section, the degrees of freedom are  $L(A-1)$ . Unfortunately, our conditions generally involve very large scores and frequently few sequences; thus, the chi-squared distribution tends to give poor probability estimates. To improve our statistical significance estimates, we are approximating probabilities after offsetting and rescaling the information content so that the average information matches the average of the distribution and the maximum possible information has the correct statistical significance. This approach is based on Read and Cressie [15] who offset and stretched a log-likelihood ratio to match the average and the standard deviation of the chi-squared distribution.

On the other hand, it is the interpretation as the large-deviation rate function that has been particularly insightful for generalizing our model to account for insertions and deletions, which we discuss in the next section. For the types of probability distributions described in this paper, a large-deviation rate function  $I(f_i)$  is a function of all the frequencies and can be defined by the following formula [5]:

$$\text{Probability } (\mathcal{U} \geq I(f_i) \geq \mathcal{L}) = f(N, \mathcal{L}, \mathcal{U}) e^{-N\mathcal{L}}, \quad (3)$$

where  $\mathcal{L}$  is any value attainable by  $I(f_i)$ ,  $\mathcal{U}$  is any value greater than or equal to  $\mathcal{L}$ , and  $f(N, \mathcal{L}, \mathcal{U})$  is a function that varies slowly relative to  $N$  and the exponential,  $e^{-N\mathcal{L}}$ , such that

$$\lim_{N \rightarrow \infty} \frac{\ln f(N, \mathcal{L}, \mathcal{U})}{N} = 0.$$

	A	A	T	T	G	A
	A	G	-	-	G	T
	A	G	-	-	G	A
	A	G	G	C	G	T
	1	2	3	4	5	6
A	4	1	0	0	0	2
C	0	0	0	1	0	0
G	0	3	1	0	4	0
T	0	0	1	1	0	2
-	0	0	2	2	0	0

Figure 2: An example of a matrix summarizing a DNA sequence alignment containing gaps. On the top is an alignment of four sequences: the first and last are 6-mers, and the middle two are 4-mers. Below is a matrix whose first four rows contain the number of times— $n_{i,j}$ —that the indicated letter is observed at the indicated position of this alignment, and whose last row contains the number of times— $n_{-,j}$ —that a gap is observed at the indicated position.

For simplicity of notation, we will indicate functions such as  $I$  without explicitly indicating the dependence on the frequencies  $f_i$ .

## 2.2 The Information Content for a Sequence Alignment Containing Gaps

In this section we describe an alignment model which contains gaps and in which each position of the alignment is considered to be independent. To develop an information formula for an alignment containing gaps, the set of aligned words is first summarized in an  $(A+1) \times L$  matrix (e.g., Figure 2). This matrix differs from the one described in section 2.1 in that it contains an additional row whose elements,  $n_{-,j}$ , are the number of times that a gap occurs at position  $j$ . In analogy with section 2.1, we define the frequencies  $f_{-,j} = n_{-,j}/N$  and  $f_{i,j} = n_{i,j}/N$ , where  $N = n_{-,j} + \sum_{i=1}^A n_{i,j}$ .

If the gaps are ignored, the probability of such an  $(A+1) \times L$  matrix is

$$P_{\text{gap matrix}} = \prod_{j=1}^L \left[ \frac{(N - n_{-,j})!}{\prod_{i=1}^A n_{i,j}!} \prod_{i=1}^A p_i^{n_{i,j}} \right]. \quad (4)$$

However, because of the presence of gaps, there are

$$M = \prod_{j=1}^L \left[ \frac{N!}{n_{-,j}!(N - n_{-,j})!} \right] \quad (5)$$

alignments consistent with this matrix when all permutations of gaps within each column are considered. For a particular configuration of gaps, the average occurrence of a particular configuration of letters is

$$A_{\text{gap matrix}} = M P_{\text{gap matrix}}. \quad (6)$$

Large-deviation principles apply as  $N$  becomes infinitely large and, thus, probabilities such as  $P_{\text{gap matrix}}$  become infinitesimally small. Thus,  $A_{\text{gap matrix}}$  can also be considered the overall probability that at least one of these  $M$  alignments has the observed occurrences of letters found in the  $(A+1) \times L$  matrix. The large-deviation rate function for  $A_{\text{gap matrix}}$  and, thus, the information content of the corresponding sequence alignment is

$$I_{\text{gap matrix}} = \sum_{j=1}^L \left[ f_{-,j} \ln f_{-,j} + \sum_{i=1}^A f_{i,j} \ln \frac{f_{i,j}}{p_i} \right]. \quad (7)$$

This last equation—a statistical basis for scoring a sequence alignment containing gaps due to insertions and deletions—is the most important conclusion of this paper.

Notice that the formula for  $I_{\text{matrix}}$  (equation 2) can be derived from the formula for  $I_{\text{gap matrix}}$  (equation 7) by setting  $f_{-,j} = 0$  because  $\lim_{f \rightarrow 0} f \log_2 f = 0$ . Also notice that any column in which  $f_{-,j} = 1$  (i.e., the corresponding position never contains any letters) contributes nothing to the overall value of  $I_{\text{gap matrix}}$ . Thus, an infinite number of meaningless positions containing only gaps can be added to any alignment without increasing the information.

$I_{\text{gap matrix}}$  is maximized under the same conditions as  $I_{\text{matrix}}$ —i.e., when the least expected letter occurs exclusively.  $I_{\text{gap matrix}}$  equals  $-L \ln 2$  and is minimized when there is an equal probability of finding a letter or a gap at each position, and the letters occur in proportion to their *a priori* probabilities—i.e.,  $f_{-,j} = 1/2$  and  $f_{i,j} = p_i/2$  for all values of  $i$ . However, in practice, the minimum of  $I_{\text{gap matrix}}$  is 0 because one can generally pick an alignment having no gaps.

Finally, to complete our derivation of the information content of an alignment containing gaps, we need a way to calculate the statistical significance of the information score. For each column of an alignment, the number of gaps can vary from zero to  $N$ . Thus, there are  $(N + 1)^L$  possible configurations of gaps in an alignment having a width of  $L$ . These configurations vary from one that contains no gaps to one that contains only gaps. If the width  $L$  of the alignment is small relative to the length of the sequences being aligned, then each configuration essentially occurs proportionally to its corresponding value of  $M$  (equation 5). Thus, we define the overall statistical significance of  $I_{\text{gap matrix}}$  as the inverse of the sum of the *average* occurrence of a score greater than or equal to  $I_{\text{gap matrix}}$  over all the  $(N + 1)^L$  configurations due to gaps. To calculate this overall statistical significance, we consider the probability distribution of

$$P = \prod_{j=1}^L \left[ \frac{N!}{n_{-,j}! \prod_{i=1}^A n_{i,j}!} (1/2)^{n_{-,j}} \prod_{i=1}^A (p_i/2)^{n_{i,j}} \right] \quad (8)$$

and its large-deviation rate function of

$$I = I_{\text{gap matrix}} + L \ln 2 \quad (9)$$

where  $I_{\text{gap matrix}}$  is defined in equation 7. The overall statistical significance of  $I_{\text{gap matrix}}$  is equal to the inverse of the product of  $2^{NL}$  and the probability of a large-deviation rate function greater than or equal to  $(I_{\text{gap matrix}} + L \ln 2)$  based on the probability distribution in equation 8. We calculate this probability as described in section 2.1 by offsetting and rescaling the large-deviation rate function so that its average value matches the average of the distribution and its maximum possible value has the correct statistical significance.

## 2.3 Information Content Incorporating Correlations

In this section we describe a model which includes correlations between the positions of the alignment—i.e., the positions are not considered to be independent. We will limit our discussion to correlations between adjacent positions of the alignment and consider only

	A	A	T	T	G	A
	A	G	-	-	G	T
	A	G	-	-	G	A
	A	G	G	C	G	T
	1	2	3	4	5	6
A	4	1	0	0	0	2
C	0	0	0	1	0	0
G	0	3	1	0	4	0
T	0	0	1	1	0	2
-	0	0	2	2	0	0
<i>ll</i>	0	4	2	2	2	4
<i>-l</i>	0	0	0	0	2	0
<i>l-</i>	0	0	2	0	0	0
<i>--</i>	0	0	0	2	0	0

Figure 3: An example of a matrix summarizing a DNA sequence alignment containing gaps and including gap-letter correlations. On the top is an alignment of the four sequences appearing in Figure 2. Below is a matrix whose first five rows are the same as those in Figure 2. The four lower rows contain the number of times— $n_{ki,j}$ —that a letter ( $l$ ) or a gap (-) is preceded by a letter or a gap.

correlations between gaps and letters without distinguishing the individual letters. This model (Figure 3) adds four additional rows to the alignment matrix described in section 2.2. The additional rows are:  $n_{ll,j}$ , the number of occurrences of a letter at position  $j$  having followed a letter at position  $j - 1$ ;  $n_{-l,j}$ , the number of occurrences of a letter at position  $j$  having followed a gap at position  $j - 1$ ;  $n_{l-,j}$ , the number of occurrences of a gap at position  $j$  having followed a letter at position  $j - 1$ ; and  $n_{--,j}$ , the number of occurrences of a gap at position  $j$  having followed a gap at position  $j - 1$ . This model adds one additional degree of freedom to each column, but only adds  $(L - 1)$  additional degrees of freedom to the statistical model because the first column has no preceding correlations.

Analogous to the previous section, the large-deviation rate function and, thus, the information content of the corresponding sequence alignment is

$$I_{\text{cor matrix}} = I_{\text{gap matrix}} + \sum_{j=2}^L \left[ \sum_{k=l,-} \sum_{i=l,-} f_{ki,j} \ln \frac{f_{ki,j}}{f_{k,j-1} f_{i,j}} \right]. \quad (10)$$

where  $f_{ki,j} = n_{ki,j}/N$ , and  $I_{\text{gap matrix}}$  is defined in equation 7.

The correlation component on the right of equation 10 is called mutual information [7]. Mutual information is always non-negative; however, the maximum possible information is not increased over its value in the absence of correlation, because the correlation information goes to zero when  $I_{\text{gap matrix}}$  is at its maximum value—i.e., when there is no variability in one of the corresponding positions of the alignment. Statistical significance is calculated similarly as in the previous section except that  $I_{\text{cor matrix}}$  is substituted for  $I_{\text{gap matrix}}$  and the degrees of freedom increase by  $(L - 1)$ .

In biological sequence alignments, it is frequently desirable to cluster gaps in adjacent positions within an alignment rather than spreading the same number of gaps throughout that alignment. Both formulas 7 and 10 are increased by clustering gaps within fewer columns. However, only formula 10 is increased by clustering gaps into adjacent positions because the necessary information involves correlations between different positions of the

alignment. In section 4, we describe an algorithm that uses both of these two formulas for determining local alignments in unaligned sequences.

This model can easily be generalized to account for correlations between individual letters; however, such a model would add  $(L - 1)A^2$  additional degrees of freedom rather than just  $(L - 1)$ . Such an expanded correlation model can be further generalized to account for non-adjacent correlations, such as necessary for describing the secondary and tertiary structures of RNA; however, determining such alignments is more difficult.

## 2.4 Information Formulas for Finite Sequence Lengths

So far we have assumed that the letters observed at a position of an alignment do not affect the expectation at any other position the alignment. This assumption is essentially true if the *a priori* probabilities are based on the large amount of sequence observed in an organism, and is exactly true if the sequences are derived from a randomized *in vitro* synthesis. However, if the *a priori* probabilities are based on the observed frequencies in the finite data set being aligned, a hypergeometric distribution is a more accurate way of modeling an alignment rather than a multinomial distribution. Let  $N_i$  be the total occurrences of letter  $i$  in all the sequences, let  $N_o$  be the total sum of all the letters outside the region of alignment, and let  $N_{i,o}$  be the total occurrences of letter  $i$  outside the region of alignment. The remaining variables are defined as in section 2.2. The information content corresponding to the hypergeometric model of a sequence alignment—ignoring correlations—is

$$I_{\text{finite matrix}} = \sum_{i=1}^A \frac{N_{i,o}}{N} \ln \frac{N_{i,o}/N_o}{p_i} + I_{\text{gap matrix}} \quad (11)$$

$$= \sum_{i=1}^A \frac{N_{i,o}}{N} \ln \frac{N_{i,o}}{N_o} + \sum_{j=1}^L \left[ f_{-,j} \ln f_{-,j} + \sum_{i=1}^A f_{i,j} \ln f_{i,j} \right] - \sum_{i=1}^A \frac{N_i}{N} \ln p_i. \quad (12)$$

The summation in equation **11** accounts for the information outside the region of alignment. If the region of alignment is small compared to the overall length of the sequences, then this extra information will likely be close to zero and  $I_{\text{finite matrix}}$  and  $I_{\text{gap matrix}}$  will be nearly the same if the *a priori* probabilities are taken to be the observed frequencies in the data set. In the other extreme, if the alignment is a global alignment, there will be no letters outside the alignment and  $I_{\text{finite matrix}}$  and  $I_{\text{gap matrix}}$  will be exactly the same. If correlations are included,  $I_{\text{cor matrix}}$  would substitute for  $I_{\text{gap matrix}}$ .

Equation **12** emphasizes that the contribution of the observed overall frequencies—the last summation in equation **12**—does not vary as a function of the particular alignment. Thus,  $I_{\text{finite matrix}}$  is essentially the same scoring formula used by Lawrence and Reilly [13] in their expectation-maximization alignment algorithm, except that those authors excluded the constant and gap portions of equation **12**. Similarly, if the constant component is excluded,  $I_{\text{finite matrix}}$  with the addition of correlation information—i.e., the summation in equation **10**—corresponds to the hidden markov models that have recently been used for determining sequence alignments [2, 11].

Calculating the statistical significance of equation **11** is somewhat different from calculating the significance of scores derived from the multinomial distribution. The details of

the calculation vary depending on whether the alignment is global or local. It is also more difficult to determine the maximum information and its significance, quantities we use for determining how to offset and rescale the information score to better fit the chi-squared distribution. In the remainder of this paper, we limit ourselves to alignment models that are based on the multinomial distributions discussed in the previous sections.

### 3 Relationship between Information and the Occurrence of a Pattern

Explicit in our definitions of information is that the higher the information of a sequence alignment, the rarer and, presumably, the more interesting the pattern described by the alignment. The following theorems link the mathematical definitions in equations **2**, **7**, and **10** with a more intuitive sense of the information of a sequence alignment. For simplicity, we will not mention correlations in the following discussion, but identical properties are true in their presence based on identical arguments. Our initial discussion in this section will be limited to alignments not containing gaps.

**Theorem:** There are a total of  $A^L$  different sequences of length  $L$ ; however, the set of L-mers contained within an alignment generally has representatives from only a small fraction of the possible sequences. We assume that the *a priori* probability of a sequence is the product of the *a priori* probabilities of its component letters. If the sequences represented in a set of aligned L-mers occur within the set in proportion to their *a priori* probabilities,  $e^{-I_{\text{matrix}}}$  will be the upper limit to the *a priori* probability that an arbitrary L-mer has the same sequence as one of the L-mers contained in the alignment (e.g., Figure 4). Thus, the higher the information content, the lower will be this probability, and the more rarely will an arbitrary L-mer be expected to match a sequence contained in the alignment.

For each position in a sequence alignment, a letter can be considered permissible or not permissible. When the sequences represented in the alignment include every combination of permissible letters, a consensus sequence will be an exact description of the pattern described by the alignment. For example, the DNA recognition sequence of the restriction endonuclease *HincII* is GTYRAC (Y = C or T; R = A or G). The four recognition sequences described by the consensus are each recognized by the enzyme and are the only sequences recognized by the enzyme. Under these latter conditions,  $e^{-I_{\text{matrix}}}$  equals the probability that an arbitrary L-mer matches one of the sequences represented in the alignment (e.g., Figure 4A).

With some modifications of the terminology, the theorem just discussed generalizes to alignments containing gaps. There are an infinite number of possible sequences when there is no restriction on word length; hence, an arbitrary word may match more than one of the sequences represented in the alignment. For example, if the sequences “AB” and “ABC” are represented in an alignment, the word “ABC” will match both sequences. To accommodate this redundancy, we will refer to *averages* rather than to *probabilities*.

$$\begin{array}{l}
\mathbf{A} \quad \begin{array}{l} \mathbf{A} \ \mathbf{G} \ \mathbf{A} \\ \mathbf{A} \ \mathbf{G} \ \mathbf{T} \\ \mathbf{C} \ \mathbf{G} \ \mathbf{A} \\ \mathbf{C} \ \mathbf{G} \ \mathbf{T} \end{array} \quad \begin{array}{l} P_1 = e^{-4.16} \\ P_2 = e^{-4.16} \\ P_3 = e^{-4.16} \\ P_4 = e^{-4.16} \end{array}
\end{array}$$

A	0.5	0	0.5
C	0.5	0	0
G	0	1	0
T	0	0	0.5

$$e^{-I} = e^{-2.77} = P_1 + P_2 + P_3 + P_4$$

$$\begin{array}{l}
\mathbf{B} \quad \begin{array}{l} \mathbf{A} \ \mathbf{G} \ \mathbf{A} \\ \mathbf{A} \ \mathbf{G} \ \mathbf{T} \\ \mathbf{C} \ \mathbf{G} \ \mathbf{A} \end{array} \quad \begin{array}{l} P_1 = e^{-4.16} \\ P_2 = e^{-4.16} \\ P_3 = e^{-4.16} \end{array}
\end{array}$$

A	0.7	0	0.7
C	0.3	0	0
G	0	1	0
T	0	0	0.3

$$e^{-I} = e^{-2.89} > e^{-3.06} = P_1 + P_2 + P_3$$

$$\begin{array}{l}
\mathbf{C} \quad \begin{array}{l} \mathbf{A} \ \mathbf{G} \ \mathbf{A} \\ \mathbf{C} \ \mathbf{G} \ \mathbf{T} \end{array} \quad \begin{array}{l} P_1 = e^{-4.16} \\ P_2 = e^{-4.16} \end{array}
\end{array}$$

A	0.5	0	0.5
C	0.5	0	0
G	0	1	0
T	0	0	0.5

$$e^{-I} = e^{-2.77} > e^{-3.47} = P_1 + P_2$$

Figure 4: Examples of the relationship between  $e^{-I}$  and the probability that a random trimer will have the same sequence as one of the trimers summarized in a matrix. An equiprobable DNA alphabet is being used; thus, each letter will have an *a priori* probability of  $0.25 = e^{-1.39}$ , and each trimer will have an expected frequency,  $P_k$ , of  $0.25^3 = e^{-4.16}$ . Each example corresponds to a pattern that can be summarized with the consensus sequence “A/C G A/T.” The matrices contain the frequencies— $f_{i,j}$ —that the indicated letter is observed at the indicated position of the alignment. (A) An alignment of all four combinations of the permissible letters indicated in the consensus sequence. The consensus sequence and the matrix are both exact descriptions of the pattern, and  $e^{-I}$  equals the sum of the probabilities. (B) An alignment of three different sequences. Not all combinations of the permissible letters indicated in the consensus sequence and shown in A are represented in the alignment. The matrix is a more informative description of the pattern than the consensus sequence; however, there is a 19% difference between the sum of the probabilities and the upper limit indicated by the information content. (C) An alignment of two different sequences. This is an extreme example in which there is complete correlation between the first and third positions of the pattern. Both the consensus sequence and the matrix do an equally poor job of describing the pattern. There is a 2-fold difference between the sum of the probabilities and the upper limit indicated by the information content.

**Theorem generalized for gaps:** If the sequences represented in a set of aligned words occur within the set in proportion to their *a priori* probabilities,  $e^{-I_{\text{gap matrix}}}$  will be the upper limit to the *a priori* average number of ways that an arbitrary word will match the sequences of the words contained in the alignment. If no sequence represented in the alignment is a prefix of another, then  $e^{-I_{\text{gap matrix}}}$  can also be interpreted as the upper limit to the probability of a match.

When the average is  $\ll 1$ , the distinction between the average number of matches and the probability of a match will likely be very small.  $e^{-I_{\text{gap matrix}}}$  equals the average number of matches if the sequences represented in the alignment include every combination of permissible letters and gaps, and no two combinations generate the same sequence.

## 4 Algorithms for Determining the Alignment Having the Optimum Information Content

The main concern of this paper is in determining information formulas for scoring alignments, especially those containing gaps. Up to now we have shown the justifications for using equations 2, 7, and 10 for this purpose. However, our goal is also to apply these formulas to identify optimal alignments and determine consensus patterns describing a functional relationship. Thus, to determine the usefulness of these information formulas, we have developed programs that determine a local alignment of a set of sequences by trying to optimize a biased information content of the alignment. Algorithms that determine *local alignments* align sequence regions having a high localized similarity. This is in contrast to algorithms that determine *global alignments* and that align sequences from end to end. Other alignment algorithms besides the one described below could also use the same information formulas to rank different alignments.

### 4.1 Biasing the Information Score

A property of the information formulas is that they are always non-negative, when an alignment does not contain gaps. However, for our local alignment algorithm to determine the width of an alignment pattern, we need the score to be negative on average—even in the absence of gaps—so that an interesting alignment can appear as a region of positive information [18, 23]. Therefore, from each position of an alignment, we subtract two biases.

The first bias is the average information score—in the absence of gaps—expected from a collection of  $N$  letters occurring with the designated a priori probabilities, where  $N$  is the number of sequences in the alignment. This correction causes the score expected of an arbitrary alignment to equal zero. We also subtract this first bias to approximate the information content of the corresponding pattern as the number of sequences in the alignment goes to infinity [17]. In this latter context, we call this biased information content the *sample-size adjusted information content*, and we use this adjusted information when we want to apply the theorems in section 3.

The second bias subtracted from each position is some multiple of the standard deviation of the information score—in the absence of gaps—expected from a collection of  $N$  letters

occurring with the designated a priori probabilities. Subtraction of this second component is what causes the expected alignment score to be less than zero. The number of standard deviations to subtract should not be the same for all alignments. We try a range of values—such as 1, 1.5, and 2—and then compare the various alignments identified according to an estimate of their statistical significance (described in the next section) or according to empirical constraints. Our standard-deviation bias, which is a simple multiple of the alignment’s width, should not be confused with the standard deviation of the information content, which is a multiple of the square root of the alignment’s width. This difference partially accounts for why we must try a range of standard-deviation biases.

## 4.2 Determining Statistical Significance

Our ultimate goal is to find the alignment with the greatest statistical significance or, otherwise, satisfies known constraints. We have discussed throughout section 2 how we estimate the statistical significance of an individual alignment. However, to obtain the ultimate statistical significance, the significance for an individual alignment needs to be divided by the huge number of possible alignments that are generally available because of the different starting points available within each sequence for initiating the alignment. For example, if we have aligned 10 sequences, each 100 bases long, there are at most  $100^{10}$  possible combinations having a single contribution from each sequence. On the other hand, there are at most  $1000!/(1000 - N)!/N!$  combinations containing exactly  $N$  sequences when each sequence may contribute zero or more times to the alignment. These approximations should be sufficient as long as the width of the alignment is small relative to the overall length of each sequence.

## 4.3 The alignment algorithms

We previously described a greedy alignment algorithm that did not permit gaps and required that the user set the width of the expected alignment [10, 21]. This original algorithm was dependent on the order with which the sequences were presented to the program. We have since modified this original algorithm so that it is essentially order independent. In this section, we describe two related algorithms that determine the width of the alignment while trying to maximize the biased information content of an alignment containing a single contribution from each sequence. One algorithm does not permit gaps and the other does. The algorithm that permits gaps can score using either formula **7** or **10**. We will describe both these algorithms together and then indicate their significant differences. As with all common multiple alignment algorithms, our approach is a compromise between the need to keep the alignment algorithms computationally practical and the desire to obtain the mathematically optimum alignment.

Initially, all possible pairwise alignments of the individual sequences are determined so as to maximize the biased information content. The pairwise alignments are ranked according to their biased information content, and some user-determined number of the highest scoring alignments are saved. Each saved alignment is completed by aligning the sequence ends not included in the local alignment, since these end regions might become incorporated into the local alignment as additional sequences are added into the alignment.

Each alignment saved from the previous step is then paired with each sequence not already contained in the alignment, and the two are aligned to form a new alignment containing an additional sequence. The alignments are scored according to the change in the biased information content resulting from aligning the additional sequence with the previously aligned sequences. Thus, each new alignment has the maximum possible biased information content that is consistent with the previously aligned sequences being held constant relative to each other. Once again, the user-determined number of alignments having the highest biased information content is saved, and the excluded ends of these saved alignments are aligned. Whenever two alignments appear to be identical, only one of the two alignments is saved. The procedure described in the current paragraph is repeated until each sequence has been incorporated into a final alignment.

There are three major differences between the algorithm that allows gaps and the algorithm that does not. First, the alignments that can have gaps are determined using dynamic programming [18]. The alignments that do not permit gaps are determined more directly. Second, when gaps are permitted, only the highest scoring alignment is determined during the alignment procedure. When gaps are not permitted, all the possible alignments are determined and considered for saving. However, there are straightforward ways to determine sub-optimal alignments even when gaps are permitted [24] that we might incorporate in the future. Third, when gaps are permitted, each alignment containing more than two sequences is refined by removing one of the sequences from the alignment and realigning that sequence with the rest of the alignment (similar to Barton and Sternberg [3]). This refinement is sequentially repeated with each sequence in the alignment until the biased information content of the alignment stops increasing. Such a refinement could also be easily added to the alignment algorithm that does not permit gaps.

A variant of the above algorithms allows each sequence to contribute zero or more words to each alignment. Each position of each sequence is permitted to contribute to the same alignment position at most once. With this algorithm, the user needs to determine the maximum number of sequences in each alignment. However, the best alignment is generally determined by maximizing the statistical significance described in section 4.2. This criteria adjusts for the number of sequences in the alignment so that alignments containing differing numbers of sequences can be compared.

## 5 Aligning LexA Binding sites

The local alignment program that did not permit gaps was tested on a collection of DNA sequences containing binding sites for the *E. coli* LexA protein. LexA represses the transcription of the genes involved in the SOS response, an inducible system for repair of bacterial DNA damage. LexA represses transcription of the SOS genes—including itself—by binding to DNA near the RNA start site. These binding sites had previously been analyzed with our program that required the presumed width of the binding site to be supplied by the user [10].

Our analysis included 10 promoter sequences, each 200 base-pairs long, having one or more binding sites for the LexA repressor. The program was set to allow each strand of each DNA sequence to contribute zero or more binding sites. The *a priori* frequencies were

assumed to be the observed frequencies of 30% A and T and 20% of G and C, which worked better than the genomic frequencies of 25% for each base when each strand was not required to have exactly one binding site.

The goal was to identify the alignment having the greatest overall statistical significance—i.e., the smallest predicted frequency as described in section 4.2. The greatest overall significance occurred when the bias was set to either 1 or 1.5 standard deviations (section 4.1) and the pattern contained 26 binding sites. This pattern contains both strands of all 12 *lexA* binding sites that have been identified in the literature on the 10 sequences being analyzed, plus an additional site that is adjacent to a known site and was suspected in our previous analysis [10].

The optimal alignment was a 24 bp symmetrical pattern having a sample-size adjusted information content (defined in section 4.1) of 13.8, indicating the pattern occurs less than once every  $e^{13.8} = 1 \times 10^6$  bases—according to the theorem in section 3—or once every  $1 \times 10^6$  base *pairs* since the pattern is symmetrical. Sequence words contained in the alignment were scored against the pattern as described in Hertz *et al.* [10]. The score of the lowest scoring sequence was calculated [19] to occur once every  $3 \times 10^5$  bp in a genome having *a priori* probabilities of 30% A and T and 20% of G and C, or once every  $2 \times 10^6$  bp in a genome having equiprobable *a priori* probabilities like the actual *E. coli* genome. Thus, the alignment pattern should be excellent at discriminating LexA binding sites from the rest of the *E. coli* genome.

We also repeated the alignment of the ten LexA-binding sequences in the presence of a randomized version of each sequence so that there were a total of twenty 200 base pair sequences. The resulting alignment was the same as obtained without the random sequences, except that one of the random sequences contributed both strands of an additional site so that the alignment contained 28 sequences rather than 26.

## 6 Aligning *E. coli* Promoter Sequences

The local alignment program was tested on a set of *E. coli* transcriptional promoters obtained from Harley and Reynolds [9]. The standard model of the *E. coli* promoter is centered around two sequence elements, each 6 bases long, which are referred to as the  $-10$  and  $-35$  regions because of their approximate distances upstream of the transcriptional start site. The spacing between the two elements is usually  $17 \pm 1$  bases, but can range from 15 to 21 bases [9]. We aligned 110 of the 231 promoter sequences in the data set.

The *E. coli* promoter is expected to occur fairly frequently, perhaps once every 1000 base pairs, which corresponds to once every 2000 bases since the promoter is not symmetrical. Therefore, the information content of a pattern having a useful predictive ability is expected to be at least  $\ln 2000 = 7.6$ . The published alignment of these 110 sequences had a sample-size adjusted information content (defined in section 4.1) of  $\ln 312 = 5.7$  (based on equation 10) in a region extending from 10 bases upstream of the  $-35$  region through the  $-10$  region. The extent of this region was based on an inspection of where the published alignment had significant concentrations of information. Thus, the information content of the published alignment seems low relative to the expected occurrence of promoters in the *E. coli* genome.

Because the *E. coli* promoter is so degenerate, we used as much prior knowledge as possible in determining our alignment. We assumed that each sequence only contained a single promoter, and that the *a priori* nucleotide frequencies were 25% for each base, since that is the frequency observed in *E. coli*. We also forced the alignment of the transcriptional initiation sites. We excluded information from adjacent gap-letter correlations, because preliminary results indicated that including such correlations cause peculiar alignments with large regions of gaps that are not consistent with our understanding of how proteins interact with DNA. Gap-letter correlations would probably be more useful with protein alignments where loops having very variable lengths are completely reasonable and expected. Since DNA is relatively rigid over the distances we are considering, large stretches of gaps are probably not reasonable in this example. Unfortunately, when biased by 1.5 standard deviations, our program achieved essentially the same sample-size adjusted information content as the published alignment within the same region. The information content is lowered by 0.4 when the alignment is extended through the position following the transcriptional initiation site because of the need to have gaps between the  $-10$  region and the start site of transcription.

In a review of 107 bacterial promoters by Collado-Vides *et al.* [6], it was observed that 45% of the promoters are associated with upstream activators. Thus, perhaps the generic promoter is very non-specific, and the apparent greater specificity is the result of the interactions with specific activators. To avoid the complexities introduced by promoters that require activators, we aligned 18 promoters that were not activatable according to this review. When the standard deviation bias was set to 1.5, the resulting alignment pattern went from the beginning of the  $-35$  region through the second transcribed position and had a sample-size adjusted information content of 9.0. Thus, this pattern would be expected to occur once every  $e^9/2 = 4000$  base pairs—a reasonable frequency for a pattern describing 55% of the *E. coli* promoters.

As a control, we also aligned a set of 18 promoters containing 10 of the non-activatable promoters and 8 activatable promoters. As expected, the alignment of this mixed set of promoters had a smaller sample-size adjusted information content of 7.1, even though this alignment extended an additional 6 positions upstream of the  $-35$  region. However, this mixed set of 18 promoters still aligned with a higher information content than the alignment of the 110 sequences. This difference may simply be an artifact of the different sizes of the two data sets, or may reflect additional differences between the quality of the two sets.

Thus, much of the difficulty in identifying *E. coli* promoters appears to be due to the requirement for auxiliary activators for approximately half of the promoters. An alignment of all 59 non-activatable promoters listed in Collado-Vides *et al.* [6] would be useful because a set of only 18 promoters is more susceptible to chance fluctuations. For example, the width of this smaller data set was much more influenced by the value of the standard deviation bias than was the larger data set of 110 promoters.

## 7 Conclusion

Information content (equation 2) has been very useful for analyzing alignments of DNA sequences [4, 10, 20, 21]. However, one of its major drawbacks has been its failure to allow for insertions and deletions in sequence alignments. We have now extended the formula for

information content to allow for these. The formula we previously used (equation **2**) can be interpreted as a special case of the extended formulas (equations **7** and **10**). As shown in the previous sections, all these information formulas share analogous properties. They are each the large-deviation rate function for the probability of an alignment, and they are similarly related to the occurrence that a random sequence will match the pattern described by the corresponding frequency matrix (section 3). In section 2.3, we explain how more complicated models that incorporate correlations can be developed. Specifically, equation **10** incorporates adjacent correlation between gaps and letters so that the information is increased by clustering gaps into adjacent columns.

We have developed algorithms that uses the various information formulas (equations **2**, **7**, and **10**) to determine regions of local alignment and, thus, derive biologically meaningful patterns. We have generally applied these alignment algorithms to the identification of the DNA binding sites of proteins; however, these algorithms can also be applied to protein motifs and RNA sequences. In this paper, we apply our techniques to identify the binding sites of the *E. coli* repressor protein LexA and patterns describing *E. coli* promoters.

A limitation of the current information formulas is their inability to incorporate pairwise letter biases such as contained in the amino-acid mutational distance matrix of Dayhoff *et al.* [8]. This limitation has restricted the value of our programs in aligning protein sequences, although this might not be a problem with a large data set (for example see Krogh *et al.* [11]). We are considering methods for adapting the information formulas to account for pairwise biases.

Past scoring schemes for comparing sequence alignments were based on intuitive ideas of how to penalize gaps. The formulas presented in this paper (equations **7** and **10**) represent an alternative approach in which the penalty is a function of the specific alignment. We are currently using this formula to develop algorithms for aligning multiple, functionally related, nucleic-acid and protein sequences. Current results indicate that this scoring scheme is indeed successful in comparing different multiple alignments. We expect the scoring schemes and formalisms presented in this paper to have a wide application to other alignment algorithms.

## Acknowledgements

We wish to thank Calvin Harley for providing us with the *E. coli* promoter sequences in a computer-readable form, and Julio Collado-Vides for helpful comments on the analysis of the *E. coli* promoter sequences. We also wish to thank Mark Borodovsky for helpful comments on this manuscript.

This work was supported by Public Health Service grants HG-00249 from the National Institutes of Health.

## References

- [1] D. J. Bacon and W. F. Anderson. Multiple sequence alignment. *J. Mol. Biol.*, 191:153–161, 1986.

- [2] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA*, 91:1059–1063, 1994.
- [3] G. J. Barton and M. J. E. Sternberg. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, 198:327–337, 1987.
- [4] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750, 1987.
- [5] J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. John Wiley and Sons, Inc., New York, 1990.
- [6] J. Collado-Vides, B. Magasanik, and J. D. Gralla. Control site location and transcriptional regulation in *escherichia coli*. *Microbiol. Rev.*, 55:371–394, 1991.
- [7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, 1991.
- [8] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure, Volume 5, Supplement 3*, chapter 22, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.
- [9] C. B. Harley and R. P. Reynolds. Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.*, 15:2343–2361, 1987.
- [10] G. Z. Hertz, G. W. Hartzell III, and G. D. Stormo. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, 6:81–92, 1990.
- [11] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, 235:1501–1531, 1994.
- [12] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.
- [13] C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41–51, 1990.
- [14] M. Murata, J. S. Richardson, and J. L. Sussmann. Simultaneous comparison of three protein sequences. *Proc. Natl. Acad. Sci. USA*, 82:3073–3077, 1985.
- [15] T. R. C. Read and N. A. C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*, chapter 5. Springer-Verlag, New York, 1988.

- [16] D. Sankoff and R. J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 253–263. Addison-Wesley, Reading, MA, 1983.
- [17] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986.
- [18] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.
- [19] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.*, 5:89–96, 1989.
- [20] G. D. Stormo. Consensus patterns in DNA. In R. F. Doolittle, editor, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, volume 183 of *Methods in Enzymology*, pages 211–221. Academic Press, San Diego, CA, 1990.
- [21] G. D. Stormo and G. W. Hartzell III. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, 86:1183–1187, 1989.
- [22] G. D. Stormo and M. Yoshioka. Specificity of the Mnt protein determined by binding to randomized operators. *Proc. Natl. Acad. Sci. USA*, 88:5699–5703, 1991.
- [23] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice: Review of concepts, case studies and implications. *J. Mol. Biol.*, 235:1–12, 1994.
- [24] M. S. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA-tRNA comparisons. *J. Mol. Biol.*, 197:723–728, 1987.