# AN INFORMATION THEORETIC VIEW OF GAPPED AND OTHER ALIGNMENTS

Jeanette P. Schmidt
*Dept. of Comp. and Inf. Science, Polytechnic University, 6 MetroTech, Brooklyn, NY 11201.*
*Current address: Incyte Pharmac., inc. 3174 Porter Drive, Palo Alto, CA 94304.*

We use an information theoretical framework to estimate the probability of the score of gapped alignments. With appropriate scaling, the score of a global (and with some adjustments also the score of a local) alignment of two sequences can be viewed as the difference in the number of bits needed to transmit the two sequences $T_1$ and $T_2$ under two different encoding schemes $C_1$ and $C_2$. $C_1$ is an idealized scheme, assumed to achieve an optimal encoding with respect to a distribution $p$, and the assumption that $T_1$ and $T_2$ are independent. $C_2$ is an alternate scheme, that will transmit $T_1$ and $T_2$ while taking advantage of the optimal alignment between the two. That is under $C_1$, the strings $T_1$ and $T_2$ (with respective probabilities $p(T_1)$ and $p(T_2)$), are assumed to be encoded using $C_1(T_1, T_2) = \log\left(\frac{1}{p(T_1)\times p(T_2)}\right)$ bits. By slightly modifying a known Theorem we show that the probability (under $p$) that two independent sequences $T_1, T_2$ can be transmitted with an alternate encoding scheme ($C_2$) with no more than $C_1(T_1, T_2) - r$ bits is bounded by $2^{-r}$. We then show how to use this bound to derive upper bounds for the probability of gapped alignment scores between two sequences.

## 1 Introduction

Amino acid substitution matrices can be interpreted from an information theoretic perspective in a very intuitive way, as shown in the seminal paper by Altschul[1]. We follow up on this characterization and show that this framework can be used directly to obtain upper bounds on the probability of obtaining high alignment scores between two sequences. The method extends to scoring matrices that allow for insertions, deletions and affine gaps. In particular, we point out that (under appropriate scaling[11]), we can view a scoring matrix as a comparison between two encoding schemes. The score obtained is proportional to the number of bits saved when transmitting the (aligned portion of the) sequences together, as opposed to transmitting each sequence independently. A similar view was taken by Allison *et al.*[5] in the context of global alignments of DNA strings. Their aim was to find a model that would maximize the savings, i.e. a model that results in a MML (minimum message length) when transmitting the aligned sequences. We are interested in estimating the probability that the encoding suggested by the scoring matrix (and the corresponding optimal local alignment) results in a message that is r bits shorter than the message

obtained when encoding each string separately, even though the strings are unrelated. We use an elementary Theorem from Cover and Thomas [7] to show that under the assumption that the strings are unrelated, the probability of saving $r$ bits by using the encoding suggested by the scoring matrix, is bounded by $2^{-r}$. While rigorous and tight estimates for the probability of obtaining a given score under an ungapped optimal alignment have been obtained in [11,9], the proof techniques used are quite intricate and are unlikely to generalize to gapped alignments.

The upper bounds we derive, although slightly higher than the tight bounds [11,9], are based on elementary methods and can be used to upper bound the probability of alignment scores which allow affine gap penalties. The statistics of alignment scores which allow gaps has been discussed, but not resolved, in [4,2,12]. It has been shown[4] that there is a phase transition, which depends on the penalty given to gaps. For low penalties, alignment scores tend to grow linearly in the length of the sequence, while for high (or infinite) penalties they exhibit logarithmic growth. Under appropriate assumptions our results imply specific upper bounds for the probability of such alignment scores and are consistent with their analysis.

We proceed to give some background on alignments under general scoring matrices. Given two protein sequences $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_m$, one frequently seeks a subsequence $I$ in $X$ and a subsequence $J$ in $Y$, for which the score $S(I, J)$ is maximal over all possible choices of $I$ and $J$. The score $S(I, J)$ is computed by aligning $I$ and $J$, i.e. by pairing symbols in $I$ with symbols in $J$, subject to the restriction that if lines were drawn between paired symbols, the lines would not cross. Whenever the paired symbols correspond to amino acid $i$ and $j$ respectively, a score $s_{ij}$ is assigned to the pair. The set of scores $s_{ij}$ constitute the scoring matrix. A continuous region of size $k$ in $I$ (resp. $J$) that is not paired with any symbol in $J$ (resp. $I$) is typically assigned a score of $-a-(k-1)b$, for $a, b > 0$. The score $S(I, J)$ is the sum of the scores for all aligned symbols and non-aligned regions. Given $I$ and $J$, $S(I, J)$ is easily computable by dynamic programming, and so is $M(X, Y) = max_{I,J} S(I, J)$ [16], as well as the *local* alignment $A(I, J)$, which corresponds to the maximal score. A central question, addressed among others in [4,11,9] concerns the significance, or probability of obtaining a score $M(X, Y)$. In most cases the probability of obtaining a given score by chance is computed under the assumption that the sequences $X$ and $Y$ are made up of independent identically distributed (i.i.d) symbols $X_1, \dots X_n$, from some alphabet $\Sigma$, ($\Sigma$ being the set of amino acids in the case of proteins). Character $i$, (amino acid $i$) is assumed to occur with probability $p(i) = p_i$ in $X$ and $p'(i) = p'_i$ in $Y$, (frequently one assumes $p = p'$.). If the probability that $M(X, Y) > r$ under these assumptions is

below a predetermined threshold, and we have obtained such a score, we will tend to conclude that $X$ and $Y$ are probably not independent, but are likely to correspond to related proteins, and vice versa. Scoring matrices and alignments also have an interesting information theoretic interpretation [1], which is the basis of our approach, and which we describe in Section 2. In Section 3 we show how to obtain probability estimates through this formulation. In Section 4 we extend the methods to gapped alignments and compare our analytical bounds to published experimental results.

## 2 Scoring Matrices and Alignments

The distribution of ungapped alignments is well understood and has been extensively analyzed [1,9,11]. Although the bounds for ungapped alignments derived through our approach are not as tight as the known bounds, we first describe our approach for this case. The simplicity of our approach will allow its generalization to gapped alignments and alignments of repetitive elements.

Given two sequences $X$ and $Y$ with i.i.d symbols (amino acids) with distributions $p$ and $p'$ respectively, we compute $M(X,Y) = max_{I,J} S(I,J)$, over all subsequences $I \in X$ and $J \in Y$. When restricting our attention to local alignments without gaps, (i.e. for the case that all symbols in $I$ are paired with some symbol in $J$ and hence $|I| = |J|$), Karlin and Altschul [11] show that under appropriate scaling any set of scores $s_{ij}$ that satisfy two simple conditions can be interpreted as log ratios of probabilities. They point out that provided that at least one $s_{ij}$ is positive and the expected score is negative, i.e. $\sum_{i=1}^{20}\sum_{j=1}^{20} p_i p'_j s_{ij} < 0$, the equation

$$\sum_{i=1}^{20}\sum_{j=1}^{20} p_i p'_j e^{\lambda s_{ij}} = 1$$

has a unique positive solution. For simplicity of notation in later sections we prefer to compute $\lambda_2 = \lambda / \ln 2$, so that $\sum_{i=1}^{20}\sum_{j=1}^{20} p_i p'_j 2^{\lambda_2 s_{ij}} = 1$. Setting $q_{ij} = p_i p'_j 2^{\lambda_2 s_{ij}}$, we obtain that $\sum_{i=1}^{20}\sum_{j=1}^{20} q_{ij} = 1$, and $q_{ij}$ can be interpreted as the target frequency of pairing $i$ and $j$ in a local alignment of $X$ and $Y$. Scaling [1] the scores of the matrix by the factor $\lambda_2$ and rewriting them in terms of these target frequencies gives $\lambda_2 s_{ij} = \log_2\left(\frac{q_{ij}}{p_i p'_j}\right)$. Note that multiplying all the values of the matrix by a constant $\lambda$ simply multiplies all alignment scores by $\lambda$ and does hence not affect the scoring scheme. Note however that adding a constant factor to the values, as in [6], while applicable to global alignments, completely changes the scoring scheme in the context of local alignments.

**The observations in the remainder of this and in the next Section form the basis for our analysis.** From a minimal length encoding perspective, consider, as was done in [5] transmitting the strings $X$ and $Y$ under two different encoding schemes and recording the difference in encoding lengths. $C_1$ is an encoding scheme that uses the (assumed) distribution of $X$ and $Y$ and transmits character $i$ in $X$ with $\log_2 \frac{1}{p_i}$ bits and character $i$ in $Y$ with $\log_2 \frac{1}{p'_i}$ bits. As long as $\sum p_i$ and $\sum p'_i$ sum to 1, such a code can be achieved via Huffman coding (up to rounding errors) or by arithmetic coding, which essentially allows us to ignore rounding errors [17]. After having computed the optimal alignment $A(X,Y)$ with score $M(X,Y)$, we consider the following alternate encoding $C_2$, (assuming for now that the aligned portions of $X$ and $Y$ contain no gaps). In $C_2$ we transmit the unaligned portions of $X$ and $Y$ as before, but transmit the aligned portions "together". In particular the pair of characters $(i,j)$ would be encoded with $\log_2 \frac{1}{q_{ij}}$ bits, (feasible since $\sum q_{ij} = 1$). If $I = i_1 \ldots i_l$ and $J = j_1 \ldots j_l$ are the aligned subsequences of $X$ and $Y$, then transmitting these aligned portions using $C_2$ would result in an encoding that has

$$\sum_{i=1}^{l} \log_2 \frac{1}{p_{i_l}} + \log_2 \frac{1}{p'_{j_l}} - \log_2 \frac{1}{q_{i_l j_l}} = \sum \lambda_2 s_{i_l j_l} = \lambda_2 S(I,J) \text{ fewer bits .}$$

To use $C_2$ however for the aligned portions, we will also have to specify where the $C_2$ encoded portion is to be inserted in the respective protein sequences. This would require roughly an additional $\log_2 n + \log_2 m$ bits, (assuming $|X| = n$ and $|Y| = m$). The difference $(\lambda_2 S(I,J) - \log_2 nm)$ constitutes the actual savings, as summarized in the following Lemma.

**Lemma 1** *Given two strings $X$ and $Y$, of lengths $m$ and $n$, assume that the score $M(X,Y)$ under a scoring matrix with parameter $\lambda_2$ is $r$. Let $C_1$ be a code that transmits an occurrence of character $i$ in string in $X$ (resp. $Y$) with $\log_2 \frac{1}{p_i}$ bits (resp. $\log_2 \frac{1}{p'_i}$ bits. Let $C_1(X,Y)$ be the number of bits needed to transmit $X$ and $Y$ in this manner. The code $(C_1, C_2)$ (as previousely described) transmits strings $X$ and $Y$ in $C_1(X,Y) - (\lambda_2 M(X,Y) - \log_2(m * n))$ bits.*

□

What is the probability that two independent strings $X$ and $Y$ with respective distributions $p$ and $p'$ can be transmitted using code $(C_1, C_2)$ with $r$ fewer bits than suggested by "the optimal encoding $C_1$"? As shown above, an upper bound for the probability of the above (encoding) problem directly leads to an upper bound for the probability that the score of a local alignment exceeds a certain value. To estimate the probability that using $C_2$ for the

aligned portions, provides a more efficient encoding, we use a Theorem from Cover and Thomas [7].

## 3 Codes and Probabilities

**Theorem 2 [Cover and Thomas, 1991; Competitive Optimality of Shannon Code]** *Let $\ell(X)$ be the codeword length associate with the Shannon code, that is, $X$ is encoded in $\ell(X) = \lceil \log_2 \frac{1}{p(X)} \rceil$ bits, and let $\ell'(x)$ be the codeword length associated with any other code. Then*

$$Prob\{\ell(X) \geq \ell'(X) + c\} \leq 2^{-c+1}$$

□

We will use the following Corollary of the above Theorem

**Corollary 3** *Let $\ell^*(X) = \log_2 \frac{1}{p(X)}$, (i.e. not necessarily corresponding to an integral number of bits), and let $\ell'(x)$ be the codeword length associated with any other code. Then $Prob\{\ell'(X) \leq \ell^*(X) - c\} \leq 2^{-c}$.*

We now apply this Corollary to our previous setting.

**Theorem 4** *Let $X$ and $Y$ be two protein sequences of length $m$ and $n$, with respective distribution $p$ and $p'$. Let $M$ be a scoring matrix with parameter $\lambda_2$ as defined earlier [11], and let $M(X,Y)$ be the score of the highest scoring local alignment of $X$ and $Y$. The probability that $M(X,Y) > r$ is bounded by $2^{-\lambda_2 M(X,Y) + \log_2(mn)}$, ($\lambda_2$ as defined in Section 2).*

*Proof:*

It follows from Lemma 1 that $M(X,Y) > r$ implies that the strings $XY$ can be transmitted using code $(C_1, C_2)$ in $(\lambda_2 M(X,Y) - \log_2(m * n))$ fewer bits than using code $C_1$ alone. By Corollary 3 the probability of this event is bounded by $2^{-\lambda_2 M(X,Y) + \log_2(mn)}$. □

The obvious question that arises is "how tight is the above bound?" The bounds derived in [11,9] imply that

$$Prob\{M(X,Y) > r\} <= 1 - e^{-e^{-\lambda M(X,Y) + \ln(Kmn)}},$$

for a constant $K$ which depends on the probabilities $p$, $p'$ and the scoring matrix. Using the inequality $1 - e^{-x} \leq x$, (and $1 - e^{-x} \approx x$ for small $|x|$) this is closely estimated [11] by $Ke^{-\lambda M(X,Y) + \ln(mn)} = K \times 2^{-\lambda_2 M(X,Y) + \log_2(mn)}$.

This shows that at least for the gapless case our approach leads to probability estimates that are close to the tight bounds, (to within the factor $K$). The simplicity of the approach allows us to get estimates for the significance of more complicated alignments, such as gapped alignments and alignments of repetitive elements.

## 4 Scores for alignments which allow gaps

We now show how to extend the estimates to the most interesting alignments, namely alignments which allow gaps. One difficulty, in our context, is to translate the score $M(X,Y)$ of a gapped alignment to a comparison between two encoding schemes. We initially follow a similar approach as [6] and then describe encodings that make use of the fact that the encoded alignments are optimal.

The scheme $C_1$ is obviously unchanged, but a proper definition of $C_2$ requires some care. Suppose that we assign, as is customary, a score of $-a-(k-1)b$, (for $a,b>0$) to a gap of size $k$. If the encoding of the alignment is "efficient", after determining the scale factor $\lambda_2^{(g)}$, $\alpha = 2^{-\lambda_2^{(g)} \times a}$ should correspond to the probability of the occurrence of a gap in the optimal alignment and $\beta = 2^{-\lambda_2^{(g)} \times b}$ to the probability of extending the gap. More accurately $\alpha\beta^{k-1}$ should correspond to the probability of the occurrence of a gap of size $k$, and $\frac{\alpha}{1-\beta}$ should correspond to the probability of a gap (of any size). Although it is unlikely that gap sizes actually follow a geometric distribution, the use of affine gap penalties seems to model such behavior. In either case, whether or not these probabilities actually correspond to desired "target probabilities" these frequencies can be used to view the scores of gapped alignments as quantities which correspond to the difference of two encoding schemes. We hence assume that the occurrence of a gap of length $k$ is encoded in $-\lambda_2^{(g)}(a-(k-1)b)$ bits, while the unaligned $k$ amino acids are encoded using $C_1$. The actual local "loss" when encoding a gap is therefore $-\lambda_2^{(g)}(a-(k-1)b)$. Accordingly, given a scoring system $s_{ij}$ with gap parameters $a,b$ we seek a $\lambda_2^{(g)} > 0$, for which

$$f(\lambda_2^{(g)}) = \sum_{ij} q'_{ij} + \frac{2^{-\lambda_2^{(g)} a}}{1 - 2^{-\lambda_2^{(g)} b}} = 1, \text{ with } q'_{ij} = p_i p'_i 2^{\lambda_2^{(g)} s_{ij}}.$$

If such a $\lambda_2^{(g)}$ exists, the quantities $q'_{ij}$ and $\alpha\beta^{k-1}$ can be used as the frequencies that govern the encoding of the aligned portion of the two sequences. The scores of gapped alignments then correspond to potential savings in encodings and Corollary 3 can be applied to these parameters. Note that the change from $\lambda_2$ to $\lambda_2^{(g)}$, changes the interpretation of the target frequencies from $q_{ij}$ to $q'_{ij}$, which while counter-intuitive, appears to be a natural side-effect when substitution matrices constructed for gap less alignments are used for alignments which allow gaps. It is encouraging to note that the condition $\sum_{ij} p_i p'_j s_{ij} < 0$ alone is not sufficient to guarantee a solution for $f(\lambda_2^{(g)}) = 1$, but that parameters $a$ and $b$ must be to sufficiently large. Indeed, $f(0) = \infty$, (for any

constant $b$), and $f$ first decreases and then increases. It follows that $f(\lambda) = 1$ has at least one solution if and only if $min_\lambda f(\lambda) <= 1$. (Two solutions exist when $min_\lambda f(\lambda) < 1$ in which case the larger one is used as a scaling factor.) A consequence of the non-existence of a solution is that there is no obvious encoding $C_2$ corresponding to the score $M(X,Y)$ and therefore Corollary 3 can not be applied. Indeed, it was shown [4] that if gaps are permitted with to small a penalty, the expected score $M(X,Y)$ grows linearly in $|X|$ resulting in a scoring scheme unsuitable for distinguishing random events from "rare and significant events". The non-existence of a solution for $f(\lambda) = 1$ might therefore be interpreted as parameters $a, b$ that are either in (or close to) the linear range. The following Lemma is a consequence of Corollary 3 and the above discussion.

**Lemma 5** *Given a scoring system $s_{ij}$ and gap parameters $a, b$, let $\lambda = \lambda_2^{(g)}$ be the largest positive solution, (if such a solution exists) for the equation*

$$\sum_{ij} p_i p'_j 2^{\lambda s_{i,j}} + \frac{2^{-\lambda a}}{1 - 2^{-\lambda b}} = 1.$$

*If such $\lambda_2^{(g)}$ exists the probability of obtaining a gapped alignment score of at least $r$ when aligning two sequences of respective lengths $m$ and $n$ is bounded by $2^{-\lambda_2^{(g)} r + \log_2(mn)}$.*

□

We computed values for $\lambda^{(g)} = \lambda_2^{(g)} \ln 2$, for the matrices BLOSUM50, BLOSUM62 and PAM250, frequencies $p' = p$ (as given in [14]) and various values for $a$ and $b$. These values for $\lambda$ were considerably lower than the stochastic estimates inferred by Altshul and Gish[2]. While they warn[2] that their (stochastically estimated) values for $\lambda$ may be an overestimate of the true asymptotic values, we proceed to show that the estimated values based on Lemma 5 are indeed to low. Note nevertheless, that our methods **guarantee** an analytically easily computable **upper bound** for the probability of obtaining a given score. We now examine improved encodings to get tighter estimates.

## 5 Improved Encodings

Before we embark on the description of more subtle encodings, we note that the only purpose of the encodings is a better understanding of the effect of allowing gaps in alignments and for the estimation of the statistical significance of such alignments. Clearly we are not proposing to use scoring matrices as guidelines for encoding or compressing sequences. One obvious shortcoming of the encoding proposed in the previous section is the fact that we "reserve" a

number of bits for the encoding of gaps when in fact gaps cannot occur in many locations in the alignment. In particular when the first $x$ aligned characters have a score that is below $a+(k-1)b$ **we know** that no gap of size $k$ or larger could possibly extend the current alignment, as the score would drop below 0. In addition, after introducing a gap and hence reducing the score by, say $g$, we know that the score must increase by at least $g$ in the remainder of the alignment. This is most striking when the alignment has a single gap. The score must be at least $g$ before a gap with penalty $g$ can be inserted and must increase by at least $g$ after the gap. If we make the reasonable assumption that in an optimal aligment after the insertion of a gap with penalty $g$ the score of the alignment must increase by at least $g$ before another gap can be inserted, it is easy to improve the encoding based on the optimal alignment, resulting in a net effect that can be viewed as an increase in the gap penalty, as shown below.

**Theorem 6** *Assume that we have a scoring matrix $s_{ij}$, and let $\lambda_2{}^{11}$ (as before) be the unique positive solution to $\sum_{i,j} p_i p_j 2^{\lambda_2 s_{ij}} = 1$. Suppose that we use gap penalty parameters $a, b$, (i.e. the score for a gap of size $k$ is $-a-(k-1)b$). If there exists a $\lambda_2^{(g)} > 0$, so that*

$$f(\lambda_2^{(g)}) = \sum_{i,j} p_i p_j 2^{\lambda_2^{(g)} s_{ij}} = 1 - \frac{2^{-\lambda_2 a}}{1-2^{-\lambda_2 b}},$$

*then an alignment score of $r$ can be interpreted as a savings of $\lambda_2^{(g)} r$ bits when encoding the aligned portion of the sequences and Corollary 3 can be applied. It follows that the probability that the highest scoring gapped alignment between two random sequences (of respective lengths $m$ and $n$ and both with amino acid distribution $p$) exceeds $r$ is bounded by $2^{-\lambda_2^{(g)} r + \log_2(mn)}$.*

*Proof:*We first describe the encoding. The encoding will normally use frequencies $q'_{ij}$ to encode the pair of amino acids $i, j$. (We will discuss the proper encoding of gaps a bit later.) Immediately following a gap encoded in say $g$ bits we "go back" to a gapless encoding using frequencies $q_{ij}$ until the score of the alignment has increased by at least $x$, at which point we revert to an encoding based on frequencies $q'_{ij} < q_{ij}$ to "make room for the possibility of encoding a new gap". Since scoring matrices do not change the score for aligned symbols depending of whether they occur immediately after a gap or not, we can instead "change the score for a gap" to reflect the "savings" that occurs after the gap. In particular, suppose that the scoring scheme records a penalty of $\lambda_2^{(g)} x$ for a gap. Let $t_{ij} = \log_2\left(\frac{q_{ij}}{p_i p_j}\right) = \lambda_2 s_{ij}$ and $t'_{ij} = \log_2\left(\frac{q'_{ij}}{p_i p_j}\right) = \lambda_2^{(g)} s_{ij}$. The alignment following the gap recorded a savings of $\sum_{\ell=1}^{r} t'_{i_\ell j_\ell}$ for the $r$

Table 1: Analytically derived values for $\lambda^{(g)}$ for BLOSUM62 and various gap penalties $a, b$ using Theorem 5

| | BLOSUM62; $\lambda^{\infty}$=0.318, $\lambda^{min}$=.17 | | | |
|---|---|---|---|---|
| | $b=1$ | $b=2$ | $b=3$ | $b=4$ |
| $a=12$ | 0.199 | 0.271 | 0.284 | 0.290 |
| $a=11$ | undef | 0.244 | 0.268 | 0.277 |
| $a=10$ | undef | undef | 0.238 | 0.256 |
| $a=9$ | undef | undef | undef | 0.206 |

aligned characters after the gap. (Assume for simplicity that we can chose $r$ so that $\sum_{\ell=1}^{r} t'_{i_\ell j_\ell}$ equals exactly $\lambda_2^{(g)} x$.) The savings that should have been reported is $\sum_{\ell=1}^{r} t_{i_\ell j_\ell}$, and the difference between the two is used to reduce the gap penalty. A recorded gap penalty of $\lambda_2^{(g)} x$ hence corresponds to encoding the gap in $\lambda_2^{(g)} x + (\sum_{\ell=1}^{r} t_{i_\ell j_\ell} - \sum_{\ell=1}^{r} t'_{i_\ell j_\ell}) = \lambda_2 x$ bits. In summary, we solve

$$f(\lambda_2^{(g)}) = \sum_{i,j} p_i p_j 2^{\lambda_2^{(g)} s_{ij}} = 1 - \frac{2^{-\lambda_2 a}}{1 - 2^{-\lambda_2 b}},$$

and set $q'_{ij} = p_i p_j 2^{\lambda_2^{(g)} s_{ij}}$. We imagine an encoding based on the frequencies $q'_{ij}$, (and frequencies $q_{ij}$ following a gap) and encode a gap of length $k$ in $(\lambda_2 a + (k-1)\lambda_2 b)$ bits. This is clearly possible since $\sum_k 2^{-(\lambda_2 a + (k-1)\lambda_2 b)} + \sum_{i,j} q'_{ij} = 1$, (by definition of the probabilities $q'$). We just showed that the recorded cost for encoding the gap would be $(\lambda_2^{(g)} a + (k-1)\lambda_2^{(g)} b)$, (while the actual cost is $(\lambda_2 a + (k-1)\lambda_2 b)$) the former corresponding precisely to the penalty for the gap prescribed by the ($\lambda_2^{(g)}$-scaled) scoring scheme. The scoring scheme hence corresponds to an alternative encoding of the sequences and by Corollary 3 the probability of the score to exceed $r$ is bounded by $2^{-\lambda_2^{(g)} r + \log_2(mn)}$. □

The resulting values for $\lambda^{(g)} = \lambda_2^{(g)} \times \ln 2$ for the BLOSUM62 matrix and parameters are given in Table 1, which also shows the values $\lambda^{\infty}$ for ungapped alignments and the values $\lambda^{min}$, corresponding to the theoretic lower bound for $\lambda$, obtained by computing $min_\lambda \sum_{i,j} p_i p_j e^{\lambda s_{ij}}$. Reducing $\lambda$ below $\lambda^{min}$ does not allow the use of gaps with lower penalty.

The encoding suggested in Theorem 6 still does not fully explore the fact that the alignment is optimal and that its score therefore remains always above zero. In particular, no gap with penalty $g$ can be inserted before the alignment reaches a score of at least $g$. This will not have a dramatic effect when the alignment has many gaps, but does affect high scoring alignments with up to

one or just a few gaps.

If the encoding were to require that (as is the case for an alignment with only one gap) the score **on either side of the gap** increases by at least the penalty for the gap, it is easy to verify that the corresponding scaling factor $\lambda_2^{(g)}$ would solve the equation below:

$$\sum_{i,j} p_i p_j 2^{\lambda_2^{(g)} s_{ij}} + \frac{2^{-(2\lambda_2 - \lambda_2^{(g)})a}}{1 - 2^{-(2\lambda_2 - \lambda_2^{(g)})b}} = 1, \qquad A$$

The Formula derived in $A$ correspond to a reasonable model for the encodings of alignments, if the percentage of alignments with one gap or less (or even with a few gaps) is quite high. (One bit suffices to indicate if the alignment can be encoded with this restriction.) To estimate the number of alignments with at most one gap (one would expect in a random setting) we compared 5000 pairs of random sequences of length 1000 each, using the amino acid frequencies [14] and recorded the number of gaps in the highest scoring alignment. As expected this percentage was quite high with the exception of scoring schemes that were predicted to be near the linear range. The estimates given by equation $A$ are recorded in the Table 2. Numbers in italics give values corresponding to cases where the experimental data suggested that the number of alignments with one gap or less is below 60%. For many entries the percentage was above 85%. We note that while we can prove the accuracy of equation $A$ only for the case of an alignment with a single gap, the analytically computed values are in good correspondance with those estimated in [2]. In particular the tables agree almost 100% on the range of parameters $a, b$ in the significant range.

## 6 Discussion and Conclusions

We have shown that an elementary method can be used to bound the probability for the local alignment score of two sequences to exceed a value $r + \log mn$. For gapless alignments the bounds we obtain through an elementary analysis are only slightly worse than similar bounds obtained through an extremely intricate analysis. For alignments with gaps our derived values for $\lambda$, although not identical seem to follow the general pattern of the ones derived in [2]. In particular, they agree on the range of values $a, b$ that fall in the logarithmic versus linear range. Our methods can also be extended to a variety of other alignments, such as the probability of high scoring multiple approximate repeats in sequences. Since the entire analysis is encoding based it automatically can be applied to the probability of finding non-overlapping repeats above a given score in protein sequences, (the factor $\log mn$ would be replaced by $\log n^2/2$).

Table 2: Analytically derived values for $\lambda$ for BLOSUM50, BLOSUM62 and PAM250 and various gap penalties $a, b$ computed by equation $A$, as compared to those derived in Ref. 2.

| | $\lambda^\infty = .232; \lambda^{min} = .125$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | BLOSUM50, analytic | | | | BLOSUM50, stochastic [2] | | | |
| | $b=1$ | $b=2$ | $b=3$ | $b=4$ | $b=1$ | $b=2$ | $b=3$ | $b=4$ |
| $a=16$ | 0.186 | 0.201 | 0.207 | 0.210 | 0.180 | 0.207 | 0.213 | 0.222 |
| $a=15$ | *0.177* | 0.194 | 0.202 | 0.205 | 0.166 | 0.202 | 0.210 | 0.216 |
| $a=14$ | *0.166* | 0.186 | 0.195 | 0.199 | 0.140 | 0.188 | 0.201 | 0.205 |
| $a=13$ | *0.153* | *0.177* | 0.187 | 0.192 | 0.114 | 0.174 | 0.188 | 0.202 |
| $a=12$ | *0.135* | *0.164* | *0.176* | *0.183* | border* | 0.158 | 0.178 | 0.192 |
| $a=11$ | undef | *0.148* | *0.163* | *0.171* | lin.* | 0.130 | 0.167 | 0.177 |

| | $\lambda^\infty = .225; \lambda^{min} = .118$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | PAM250, analytic | | | | PAM250, stochastic [2] | | | |
| | $b=1$ | $b=2$ | $b=3$ | $b=4$ | $b=1$ | $b=2$ | $b=3$ | $b=4$ |
| $a=16$ | 0.163 | 0.181 | 0.189 | 0.193 | 0.172 | 0.200 | 0.208 | 0.217 |
| $a=15$ | *0.152* | 0.173 | 0.182 | 0.187 | 0.154 | 0.193 | 0.203 | 0.208 |
| $a=14$ | *0.138* | 0.163 | 0.173 | 0.179 | 0.131 | 0.180 | 0.194 | 0.204 |
| $a=13$ | *0.120* | *0.150* | *0.162* | 0.170 | 0.110 | 0.163 | 0.184 | 0.196 |
| $a=12$ | undef | *0.133* | *0.148* | *0.157* | border | 0.145 | 0.170 | 0.181 |
| $a=11$ | undef | undef | *0.130* | *0.140* | lin. | 0.122 | 0.153 | 0.165 |

| | $\lambda^\infty = .318; \lambda^{min} = .170$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | BLOSUM62, analytic | | | | BLOSUM62, stochastic [2] | | | |
| | $b=1$ | $b=2$ | $b=3$ | $b=4$ | $b=1$ | $b=2$ | $b=3$ | $b=4$ |
| $a=12$ | 0.275 | 0.289 | 0.295 | 0.298 | 0.275 | 0.300 | 0.305 | 0.305 |
| $a=11$ | 0.263 | 0.280 | 0.287 | 0.291 | 0.255 | 0.286 | 0.301 | 0.301 |
| $a=10$ | *0.246* | 0.268 | 0.277 | 0.281 | 0.216 | 0.266 | 0.281 | 0.293 |
| $a=9$ | *0.225* | *0.253* | 0.264 | 0.270 | 0.176 | 0.244 | 0.273 | 0.273 |

*Linear and border[2] corresponds to ranges that were judged to be in the linear range or that were borderlineas judged by the stochastic analysis.

The significance of overlapping repeats, which seems to be difficult to analyse, can also be estimated by the above method. These applications will be discussed in more detail in a full paper.

**Acknowledgements**

This work was supported through NSF VPW award HRD-9627109, and partially also supported by NSF grant CCR-9305873. We thank Eli Upfal and

Craig Nevill-Manning for several discussions on the subject.

## References


1. A.F. Altschul, *Amino Acid Substitution Matrices from an Information Theoretic Perspective*, J. Mol. Biol. 219, (1991), pp. 555-595.
2. S. F. Altschul and W. Gish, *Local Alignment Statistics*, Methods in Enzymology, (1996), 266, pp. 460–480.
3. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, *A basic Local Alignment Search Tool*, J. Molecular Biology, 215, (1990), pp. 403–410.
4. R. Arratia and M. Waterman, *A Phase Transition for the Score in Matching Random Sequences Allowing Deletions*, The Annals of Applied Probability, (1994), Vol. 4, No 1, 200-225. Algorithmica (1987) 2, pp. 195–208.
5. L. Allison, C.S. Wallace and C.N. Yee, *Finite-State Models in the Alignment of Macromolecules*, Journal of Molecular Evolution (1992) 35, pp. 77-89.
6. L. Allison, *Normalization of Affine Gap Costs Used in Optimal Sequence Alignment*, J. theor. Biol. (1993) 161, pp. 263–269.
7. T. Cover and J. Thomas, *Elements of Information Theory,* Wiley Series in Teleccommunications, John Willey & Sons, inc.
8. M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary change in proteins*, in Atlas of Protein Sequence and Structure (M.O. Dayhoff ed.) 5,3 (1978), pp. 345-352.
9. A. Dembo, S. Karlin, O. Zeitouni, *Limit Distribution of Maximal non-aligned Two Sequence Segmental Score*, The Annals of Probability, 1994, Vol 22, No. 4, 2022-2039.
10. S. Henikoff and J. G. Henikoff, *Amino acid substitution matrices from protein blocks,* Proc Natl Acad Sci U S A 89, (1992), pp. 10915–10919.
11. S. Karlin and S. F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes,* Proc. Natl. Acad. Sci. USA 87, (1990), pp. 2264–2268.
12. R. Mott, Bull. of Math. Biol. 54, (1992), pp. 59.
13. S.B. Needleman and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins,* J. of Mol. Bio., Vol. 48, (1970), pp. 443-453.
14. A.B. Robinson and L.R. Robinson, *Proc. Natl. Acad. Sci. U.S.A.* 88, (1991), pp. 8880.
15. D.Sankoff and J.B. Kruskal (editors), Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison, Addison-Wesley, Reading, MA, 1983.
16. T. F. Smith and M. S. Waterman, *Identification of common Molecular subsequences*, J. Mol. Biol., 147, (1981) pp. 195–197.
17. I. Witten, R. Neal and J. Cleary, *Arithmetic Coding for data compression*, Communications of the ACM 30, (1987) pp. 520-540.