Local Optimality Properties of Biological Sequence Alignments

Nancy R. Zhang

(Phd research with Prof. David O. Siegmund)
Biological Sequence Alignments

x: ...RNATQRNDCAMFKRRPPSPEGEHIL...
y: ...AAQDCEMFPPAPREEGDHLMLCAAT...
### Biological Sequence Alignments

**x:** \[\ldots RNATQRNDCAMFKRPPSPEGEHIL\ldots\]

**y:** \[\ldots AAQDCEMFPPAPREEGDHILMCAAT\ldots\]

**Substitution Matrix \((\bar{K})\):**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>-1</td>
<td>-2</td>
<td>\ldots</td>
</tr>
<tr>
<td>R</td>
<td>-1</td>
<td>5</td>
<td>0</td>
<td>\ldots</td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>\ldots</td>
</tr>
</tbody>
</table>

**Gap Penalties:**

- gap open: \(\Delta\)
- gap extension: \(\delta\)
Biological Sequence Alignments

\[ x: \ldots RNATQRNDCAMFKRRPPSPEGEHIL\ldots \]
\[ y: \ldots AAQDCEMFPPAPREEGDHILMCAAT\ldots \]

Substitution Matrix (\( K \)):

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>-1</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-1</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Gap Penalties: gap open: \( \Delta \), gap extension: \( \delta \)

We do not allow simultaneous gaps in both sequences.
Possible alignment: \[ z = \{(i_1, j_1), \ldots, (i_u, j_u)\} : \]

\begin{align*}
\text{x:} & \quad \ldots \text{ATQRNDCAMFKRRPPSP--EGEHIL...} \\
& \hspace{1cm} \mid \mid \mid \mid \mid \mid \mid \mid \mid \\
\text{y:} & \quad \ldots \text{AAQ--DCEMF---PPAPREEGDHIL...} \\
\end{align*}
Possible alignment: \( z = \{(i_1, j_1), \ldots, (i_u, j_u)\} \):

\[
x: \quad \ldots\text{ATQRNDCAMFKRRPPSP}--\text{EGEHIL}\ldots
\]

\[
y: \quad \ldots\text{AAQ}--\text{DCEMF}---\text{PPAPREEGDHIL}\ldots
\]

Score: \( S_z(x, y) = \sum_{k=1}^{u} K(x_{i_k}, y_{j_k}) - l\Delta - m\delta \) (here, \( l = 3 \) and \( m = 7 \)).
Possible alignment: \[ z = \{(i_1, j_1), \ldots, (i_u, j_u)\}: \]

\[
x: \ldots\text{ATQRNDCAMFKRRPPSP--EGEHIL}\ldots
\]

\[
y: \ldots\text{AAQ--DCEMF---PPAPREEGDHIL}\ldots
\]

Score: \[ S_z(x, y) = \sum_{k=1}^{u} K(x_{i_k}, y_{j_k}) - l\Delta - m\delta \] (here, \( l = 3 \) and \( m = 7 \)).

Maximum Sub-alignment Score: \[ H_n(x, y) = \max_{z \in Z} S_z(x, y) \]
Possible alignment: \[ z = \{(i_1, j_1), \ldots, (i_u, j_u)\} : \]

\[
\begin{align*}
\textbf{x}: & \quad \ldots \text{ATQRNDCAMFKRRPPSP--EGEHIL}\ldots \\
\textbf{y}: & \quad \ldots \text{AAQ--DCEMF---PPAPREEGDHIL}\ldots \\
\end{align*}
\]

Score: \[ S_z(x, y) = \sum_{k=1}^{u} K(x_{i_k}, y_{j_k}) - l\Delta - m\delta \text{ (here, } l = 3 \text{ and } m = 7) \].

Maximum Sub-alignment Score: \[ H_n(x, y) = \max_{z \in Z} S_z(x, y) \]

Null Distribution: \[ x, y \text{ iid } \sim \mu \]
Possible alignment: \( z = \{ (i_1, j_1), \ldots, (i_u, j_u) \} \):

\[
\begin{align*}
\text{x:} & \quad \ldots ATQRNDCAMFKRRPPSP--EGEHIL\ldots \\
\text{||| } \text{||| } \text{||| } \text{||| } \text{||| } \text{||| } \\
\text{y:} & \quad \ldots AAQ--DCEMF--PPAPREEGDHIL\ldots \\
\end{align*}
\]

Score: \( S_z(x, y) = \sum_{k=1}^{u} K(x_{i_k}, y_{j_k}) - l\Delta - m\delta \) (here, \( l = 3 \) and \( m = 7 \)).

Maximum Sub-alignment Score: \( H_n(x, y) = \max_{z \in Z} S_z(x, y) \)

Null Distribution: \( x, y \) iid \( \sim \mu \)

Two Questions:

How does \( H_n(x, y) \) grow with \( n \)?

What is \( P_0(H_n(x, y) > b) \) for large \( b \)?
Basic Result: The Phase Transition Phenomenon

Let $G_n(x, y)$ be the maximum alignment score of $x$ and $y$, penalizing gaps at the ends. By the theory of subadditive sequences,

$$
\alpha \overset{\Delta}{=} \alpha(K, \Delta, \delta) = \lim_{n \to \infty} \frac{\mathbb{E}(G_n)}{n} \text{ exists.}
$$
Basic Result: The Phase Transition Phenomenon

Let $G_n(x, y)$ be the maximum alignment score of $x$ and $y$, penalizing gaps at the ends.

By the theory of subadditive sequences,

$$\alpha \doteq \alpha(K, \Delta, \delta) = \lim_{n \to \infty} \frac{\mathbb{E}(G_n)}{n} \text{ exists.}$$

Arratia and Waterman (1994) Showed that

$$\begin{align*}
\alpha > 0 & \implies \mathbb{P}(\lim_{n \to \infty} \frac{H_n}{n} = \alpha) \to 1 \\
\alpha < 0 & \implies \exists \ b \ s.t. \ \forall \ \epsilon > 0, \ \mathbb{P}((1 - \epsilon)b < \frac{H_n}{\log(n)} < (2 + \epsilon)b) \to 1
\end{align*}$$
Brief Literature Review
1. Gaps NOT Allowed:

Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

$$\mathbb{E}_0[K(x, y)] < 0, \quad \mathbb{P}_0(K(x, y) > 0) > 0,$$

$H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.
Brief Literature Review

1. Gaps NOT Allowed:
   Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

   $$\mathbb{E}_0[K(x, y)] < 0, \quad \mathbb{P}_0(K(x, y) > 0) > 0,$$

   $H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.
Brief Literature Review

1. Gaps NOT Allowed:

Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

\[ \mathbb{E}_0[K(x, y)] < 0, \quad \mathbb{P}_0(K(x, y) > 0) > 0, \]

$H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.

Brief Literature Review

1. Gaps NOT Allowed:

   Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

   $$
   \mathbb{E}_0[K(x, y)] < 0, \quad \mathbb{P}_0(K(x, y) > 0) > 0,
   $$

   $H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.

   - Mott and Tribe (1999) *Journal of Computational Biology*
Brief Literature Review

1. Gaps NOT Allowed:

Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

$$\mathbb{E}_0[K(x, y)] < 0, \mathbb{P}_0(K(x, y) > 0) > 0,$$

$H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.

- Mott and Tribe (1999) *Journal of Computational Biology*
- Siegmund and Yakir (2000) *Annals of Statistics*
Brief Literature Review

1. Gaps NOT Allowed:
   Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

   $$\mathbb{E}_0[K(x, y)] < 0, \mathbb{P}_0(K(x, y) > 0) > 0,$$

   $H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.
   - Mott and Tribe (1999) *Journal of Computational Biology*
   - Siegmund and Yakir (2000) *Annals of Statistics*
   - Grossman and Yakir (2004) *Bernoulli*
Brief Literature Review

1. Gaps NOT Allowed:
   Dembo et. al. (1994, *Ann. Probab.*) showed that for scoring matrices $K$ satisfying:

   $$
   \mathbb{E}_0[K(x, y)] < 0, \quad \mathbb{P}_0(K(x, y) > 0) > 0,
   $$

   $H_n(x, y)$ grows logarithmically with $n$ and has extreme value type limiting distribution.

2. Gaps Allowed: No complete theory.
   - Mott and Tribe (1999) *Journal of Computational Biology*
   - Siegmund and Yakir (2000) *Annals of Statistics*
   - Grossman and Yakir (2004) *Bernoulli*
A Theorem from Chan (2003)

Let \((K, \Delta, \delta)\) be chosen such that the convex function

\[
h(\theta) = \left(1 + 2 \sum_{k \geq 1} e^{-\theta(\Delta + \delta k)}\right) \sum_{x, y \in A} e^{\theta K(x, y)} \mu(x) \mu(y)
\]

has a positive root of 1, with \(\tilde{\theta}\) being the larger root, then

\[
P(H_n(x, y) \geq b) \leq n^2 e^{-\tilde{\theta}b}
\]
A Theorem from Chan (2003)

Let \((K, \Delta, \delta)\) be chosen such that the convex function

\[
h(\theta) = \left(1 + 2 \sum_{k \geq 1} e^{-\theta(\Delta+\delta k)} \right) \sum_{x, y \in A} e^{\theta K(x, y)} \mu(x) \mu(y)
\]

has a positive root of 1, with \(\tilde{\theta}\) being the larger root, then

\[
P(H_n(x, y) \geq b) \leq n^2 e^{-\tilde{\theta}b}
\]

Works for \(K = \text{Blosum62}, \Delta = 18, \delta = 1\).
Let

\[ \phi(\theta) = \lim_{n \to \infty} \frac{1}{n} \log E(e^{\theta G_n}), \]

then \( \phi(\theta) = 0 \) has a positive root is necessary and sufficient for logarithmic region. The root is the large deviations rate.

A result of similar nature is given in Chan (2005).
Sketch of Proof (Chan 2003): Construct measure $Q$ on $\mathcal{A}^m \times \mathcal{A}^n \times \mathcal{Z}$ as follows:

1. Pick $(i_1, j_1)$ uniformly from $\{1, \ldots, n\}^2$, set $l = 1$.

2. Recursively, pick the aligned pair $(x_{i_l}, y_{j_l})$ from

$$f(x, y) = e^{\tilde{\theta}K(x, y) - s(\tilde{\theta})} \mu(x)\mu(y),$$

where $s(\tilde{\theta}) = log \left(1 + 2 \sum_{k \geq 1} e^{-\theta g(k)}\right)$.

and $(G^x_l, G^y_l)$, the gap at position $l$, from

$$\mathbb{P}((G^x_l, G^y_l) = (k, 0)) = \mathbb{P}((G^x_l, G^y_l) = (0, k)) = e^{-\tilde{\theta}g(k) - s(\tilde{\theta})}$$

Let $i_{l+1} = i_l + G^x_l, j_{l+1} = j_l + G^y_l$.

3. Let $z$ be the alignment produced in this process. Stop sampling when $i_l > n, j_l > n,$ or $S_z > b$. All unaligned positions are iid $\sim \mu$.

Let $Q_z$ be the measure of $(x, y)$ generated by alignment $z$. Let $Q = \sum_{z \in \mathcal{Z}} Q_z$. Let $z^*$ be the optimal alignment. Then

$$\mathbb{P}(H_n(x, y) > b) = \mathbb{E}_Q \left[ \frac{dP}{dQ}; H_n(x, y) > b \right] \leq \mathbb{E}_Q \left[ \frac{dP}{dQ_{z^*}}; H_n(x, y) > b \right] \leq n^2 e^{-\tilde{\theta}b}$$
An optimal alignment is heavily constrained around gaps...

Toy example: $\mathcal{A} = \{0, 1\}$, $K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$
An optimal alignment is heavily constrained around gaps...

Toy example: \( \mathcal{A} = \{0, 1\}, \quad K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \)

\[ x: \quad \ldots 1 \ 1 \ 0 \ 1 \ 1 \ldots \]

\[ y: \quad \ldots 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ldots \]
An optimal alignment is heavily constrained around gaps...

Toy example: \( \mathcal{A} = \{0, 1\} \), \( K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \)

\[
\begin{align*}
\text{x:} & \quad \ldots1\ 1\ 0\ 1\ 1\ldots \\
\text{y:} & \quad \ldots1\ 1\ 1\ 0\ 0\ 0\ 1\ 1\ldots \\
\text{x:} & \quad \ldots1\ 1\ 1\ 1\ 1\ldots \\
\text{y:} & \quad \ldots1\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ldots 
\end{align*}
\]
Local Optimality Property
Local Optimality Property

Definition: A Section of type \((u, v, 1)\) is \(u\) aligned letters followed by a gap of length \(v\) in x-sequence (similarly, type \((u, v, 0)\) has gap in y-sequence.)
Local Optimality Property

Definition: A Section of type \((u, v, 1)\) is \(u\) aligned letters followed by a gap of length \(v\) in x-sequence (similarly, type \((u, v, 0)\) has gap in y-sequence.

Fact: Any alignment is composed of a sequence of sections. The sum of scores of the sections is the score of the alignment. (Not including the gap penalty at the end.)
Local Optimality Property

Definition: A Section of type \((u, v, 1)\) is \(u\) aligned letters followed by a gap of length \(v\) in \(x\)-sequence (similarly, type \((u, v, 0)\) has gap in \(y\)-sequence.

Fact: Any alignment is composed of a sequence of sections. The sum of scores of the sections is the score of the alignment. (Not including the gap penalty at the end.)

Define local move: \(\phi_L\)

We can apply the move \(r\) times: \(\phi_L^r = \underbrace{\phi \ast \phi \ast \ldots \phi}_r\)
Local Optimality Property

**Definition:** A *Section* of type \((u, v, 1)\) is \(u\) aligned letters followed by a gap of length \(v\) in x-sequence (similarly, type \((u, v, 0)\) has gap in y-sequence.

**Fact:** Any alignment is composed of a sequence of sections. The sum of scores of the sections is the score of the alignment. (Not including the gap penalty at the end.)

Define local move: \(\phi_L\)

We can apply the move \(r\) times: \(\phi_L^r = \underbrace{\phi \ast \phi \ast \ldots \phi}_{r}\)

For a section \(C\) of type \((u, v, t)\),

\[
N_L(C) = |\{r : S(\phi_L^r(C)) = S(C)\}|
\]

\[
I_L(C) = \begin{cases} 
0, & \exists r \text{ s.t. } S(\phi_L^r(C)) > S(C); \\
\frac{1}{N_L(C)}, & \text{otherwise.}
\end{cases}
\]
Theorem 1

Let \((K, \Delta, \delta)\) be chosen such that the convex function

\[
2 \sum_{u \geq 1, v \geq 1} \sum_{\substack{x \in A^{u+v}, \; y \in A^u}} e^{\theta(K(x,y) - \Delta - \delta v)} I_L(x, y) \mu(x) \mu(y)
\]

has a positive root of 1, with the largest root denoted by \(\tilde{\theta}\), then exists constant \(B(\tilde{\theta})\) such that

\[
P(H_n(x, y) \geq b) \leq n^2 B(\tilde{\theta}) e^{-\tilde{\theta}b}.
\]
Sketch of Proof (1): Let
\[ q(u, v, 1) = q(u, v, 0) = \sum_{x \in A^{u+v}, y \in A^u} e^{\tilde{\theta}(K(x, y) - \Delta - \delta v)} I_L(x, y) \mu(x) \mu(y), \]
then \( q \) is a probability measure on the set of all possible section types. Also, for \( x \in A^{u+v}, y \in A^u \), let
\[ q_{u,v,0}(x, y) = q_{u,v,1}(x, y) = \frac{e^{\tilde{\theta}(K(x, y) - \Delta - \delta v)} I_L(x, y) \mu(x) \mu(y)}{q(u, v, 1)}, \]
then \( q_{u,v,\cdot} \) is a probability measure on the sequences of section type \( (u, v, \cdot) \).

Construct \( Q \) on \( A^m \times A^n \) as follows:

1. Pick \( (i_1, j_1) \) uniformly from \( \{1, \ldots, n\}^2 \), set \( l = 1 \).

2. Pick the section type \( iid \) from \( q(u, v, t) \) and the letter sequence within the section from the joint distribution \( q_{u,v,t} \).

3. Stop sampling when either the score exceeds \( b \) or one of the sequences exceeds \( n \).

Let \( Q_z \) be the measure of \((x, y)\) generated by alignment \( z \). Let \( Q = \sum_{z \in \mathcal{Z}} Q_z \).
Sketch of Proof (2): Let \( z \) have sections \( \{C_k : 1 \leq k \leq l\} \). By construction of \( Q_z \), we have:

\[
\frac{dQ_z}{dP}(x, y) = \frac{1}{n^2} \left[ \prod_{k=1}^{l} I_L(C_k) \right] b(\tilde{\theta}) e^{\tilde{\theta} S_z(x,y)}
\]
Sketch of Proof (2): Let $z$ have sections $\{C_k : 1 \leq k \leq l\}$. By construction of $Q_z$, we have:

$$\frac{dQ_z}{dP}(x, y) = \frac{1}{n^2} \left[ \prod_{k=1}^{l} I_L(C_k) \right] b(\tilde{\theta}) e^{\tilde{\theta} S_z(x, y)}$$

On the set $A = \{(x, y) : H(x, y) > b\}$, exists $z^* = z^*(x, y)$ which is locally optimal such that $S_{z^*}(x, y) > b$. Let $\Phi(z^*)$ be all alignments reachable from $z^*$ through local moves. Then,

$$\frac{dQ}{dP}(x, y) > \frac{\sum_{z \in \Phi(z^*)} dQ_z}{dP}(x, y)$$

$$> \left[ \prod_{k=1}^{l} N_L(C_k) \right] \frac{dQ_{z^*}}{dP}(x, y)$$

$$= b(\tilde{\theta}) e^{\tilde{\theta} S_{z^*}(x, y)}/n^2$$

$$> b(\tilde{\theta}) e^{\tilde{\theta} b}/n^2$$

Therefore,

$$P(A) = \mathbb{E}_Q\left(\frac{dP}{dQ}, A\right) < \frac{1}{n^2} b^{-1}(\tilde{\theta}) e^{-\tilde{\theta} b}$$
Extension to a Markov Model

We can improve on the result of Theorem 1 by also considering two adjacent sections together and allowing wobbles in the right-to-left direction across the gap, $\phi_R$.

For two adjacent sections $C_1$ and $C_2$

$$N_R(C_1, C_2) = |\{ r : S(\phi_R(C_1, C_2)) = S(C_1, C_2) \}|$$

$$N(C_1, C_2) = N_L(C_1) + N_R(C_1, C_2)$$

and

$$I(C_1, C_2) = \begin{cases} 
0, & \text{exists local move that improves the score;} \\
\frac{1}{N(C_1, C_2)}, & \text{otherwise.}
\end{cases}$$
Theorem 2

Let \( T : \mathbb{Z}_+ \times \mathbb{Z}_+ \times \{0, 1\} \to \mathbb{Z}_+ \) be any 1-1, onto map. Let \( \mathcal{M}(\theta) = \mathcal{M}(\theta, K, \Delta, \delta) \) be the matrix with elements

\[
\mathcal{M}(\theta)_{T(u_1,v_1),T(u_2,v_2)} = \sum_{x \in A^{\lceil u_1 \rceil+v_1+[u_2]+v_2}} \sum_{y \in A^{\lceil u_1 \rceil+[u_2]}} e^{\theta(K(x,y)-\Delta-\delta v_2)} I(x,y) \mu(x) \mu(y).
\]

If \((K, \Delta, \delta)\) are chosen such that there exists a value of \(\theta\) for which \(\mathcal{M}(\theta)\) has 1 as the largest eigenvalue, with \(\tilde{\theta}\) being the largest such value, then exists constant \(B(\tilde{\theta})\) such that

\[
\mathbb{P}(H_n(x, y) \geq b) \leq n^2 B(\tilde{\theta}) e^{-\tilde{\theta}b}.
\]
Lemma  Let $M$ be a matrix with positive elements. If the largest eigenvalue of $M$ is 1 and $v^L$, $v^R$ are the corresponding left and right eigenvectors, respectively, then

1. $P = D^{-1}MD$ is a stochastic matrix, where $D = \text{diag}(v^R)$.

2. Let $\pi' = [v^R_1v^L_1, v^R_2v^L_2, \ldots]$, then $\pi / ||\pi||$ is the stationary distribution of $P$.

The proof of Theorem 2 is similar to that for Theorem 1, except for the sections of an alignment are no longer drawn independently. Instead, they are drawn from a Markov Chain with transition matrix constructed from $M(\tilde{\theta})$. 
Technicalities...

Theorem 1 involves an infinite summation over all section types, and Theorem 2 involves taking the eigenvalue of an infinite dimensional matrix. In practice, we can not calculate the optimality indicator $I(\ldots)$ for all section types.
Technicalities...

Theorem 1 involves an infinite summation over all section types, and Theorem 2 involves taking the eigenvalue of an infinite dimensional matrix. In practice, we can not calculate the optimality indicator $I(\ldots)$ for all section types.

Thus we cap the number of allowable transforms to a maximum of $\kappa$. Intuitively, in most cases alignment score can not be increased by doing many consecutive transforms.
Technicalities...

Theorem 1 involves an infinite summation over all section types, and Theorem 2 involves taking the eigenvalue of an infinite dimensional matrix. In practice, we can not calculate the optimality indicator $I(\ldots)$ for all section types.

Thus we cap the number of allowable transforms to a maximum of $\kappa$. Intuitively, in most cases alignment score can not be increased by doing many consecutive transforms.

Then, the conditions for Theorems 1 and 2 can be easily verified using importance sampling based Monte Carlo.
Figure 1. For $\delta = 1$ and increasing values of $\kappa$, the minimum value of $\Delta$ that can be proven to be in the logarithmic region. Dashed line is for non-Markov result, solid line is for Markov sections result.
How well can we possibly do using local optimality?

For each $\kappa$, let $z$ be an alignment composed of two stretches of $\kappa$ aligned pairs with a gap of length $v$ in the middle. Then for all $\theta$,

$$E_z[I^\kappa(z)] = E_z \left[ \frac{\max_{\phi} e^{\theta K(\phi(z))}}{\sum_{\phi} e^{\theta K(\phi(z))}} \right]$$

Then

$$\lim_{\kappa \to \infty} E_z[I^\kappa(z)] = \lambda_v,$$

where $\lambda_v$, $v = 1, 2, \ldots$ are the constants defined in Siegmund and Yakir (2000). Storey and Siegmund (2001) showed that for all $v$, $\lambda_v \approx 0.337$. 
In effect, Theorems 1 and 2 give a new criterion function for calculating the large deviations rate. Below is a plot of the criterion functions for fixed scoring parameters $K = \text{BLOSUM62}, \Delta = 15, \delta = 1$. 
Table 1. Boundary of logarithmic region provable using Chan (2003), Theorem 1 using independent sections, and Theorem 2 using Markov sections. The last column shows numerically determined boundaries.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>Chan 2003</th>
<th>Independent Sections</th>
<th>Markov Sections</th>
<th>Altscul and Gish (1996)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.1</td>
<td>16.1</td>
<td>14.4</td>
<td>$\approx 8$</td>
</tr>
<tr>
<td>2</td>
<td>15.0</td>
<td>13.0</td>
<td>11.3</td>
<td>$\approx 6$</td>
</tr>
<tr>
<td>3</td>
<td>13.0</td>
<td>10.9</td>
<td>9.2</td>
<td>$\approx 5$</td>
</tr>
</tbody>
</table>
Final Comments
Final Comments

• Local optimality play a large role in constraining the alignment...
Final Comments

• Local optimality play a large role in constraining the alignment...

• ...but only gets us about half way there. Constraints across multiple gaps is needed to get a more precise boundary.
Final Comments

• Local optimality play a large role in constraining the alignment...

• ...but only gets us about half way there. Constraints across multiple gaps is needed to get a more precise boundary.

• We have obtained a better lower bound for the large deviations rate. How good is our lower bound remains to be investigated.
Final Comments

- Local optimality play a large role in constraining the alignment...
- ...but only gets us about half way there. Constraints across multiple gaps is needed to get a more precise boundary.
- We have obtained a better lower bound for the large deviations rate. How good is our lower bound remains to be investigated.
- These results can be generalized to other types of scoring scenarios.
Final Comments

• Local optimality play a large role in constraining the alignment...

• ...but only gets us about half way there. Constraints across multiple gaps is needed to get a more precise boundary.

• We have obtained a better lower bound for the large deviations rate. How good is our lower bound remains to be investigated.

• These results can be generalized to other types of scoring scenarios.

Thank you!