Local Optimality Properties of Biological Sequence Alignments

Nancy R. Zhang

(Phd research with Prof. David O. Siegmund)

Biological Sequence Alignments

```
x:        ...RNATQRNDCAMFKRRPPSPEGEHIL...
y:        ...AAQDCEMFPPAPREEGDHILMCAAT...
```

Biological Sequence Alignments

```
x:        ...RNATQRNDCAMFKRRPPSPEGEHIL...
y:        ...AAQDCEMFPPAPREEGDHILMCAAT...
```

Substitution Matrix (K):

	A	R	N	$\cdots$
A	4	-1	-2	$\cdots$
R	-1	5	0	$\cdots$
N	-2	0	6	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Gap Penalties: gap open: Δ, gap extension: δ

Biological Sequence Alignments

```
x:      ...RNATQRNDCAMFKRRPPSPEGEHIL...
y:      ...AAQDCEMFPPAPREEGDHILMCAAT...
```

Substitution Matrix (K):

	A	R	N	$\cdots$
A	4	-1	-2	$\cdots$
R	-1	5	0	$\cdots$
N	-2	0	6	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Gap Penalties: gap open: Δ, gap extension: δ

We do not allow simultaneous gaps in both sequences.

Possible alignment: $\mathbf{z} = \{(i_1, j_1), \ldots, (i_u, j_u)\}$:

```
x:          ...ATQRNDCAMFKRRPPSP--EGEHIL...
               |||  |||||   ||||  |||||||
y:          ...AAQ--DCEMF---PPAPREEGDHIL...
```

Possible alignment: $\mathbf{z} = \{(i_1, j_1), \ldots, (i_u, j_u)\}$:

```
x:          ...ATQRNDCAMFKRRPPSP--EGEHIL...
               |||  |||||   ||||  |||||||
y:          ...AAQ--DCEMF---PPAPREEGDHIL...
```

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

Possible alignment: $\mathbf{z} = \{(i_1, j_1), \ldots, (i_u, j_u)\}$:

```
x:      ...ATQRNDCAMFKRRPPSP--EGEHIL...
           |||  |||||   ||||  |||||||
y:      ...AAQ--DCEMF---PPAPREEGDHIL...
```

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

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

Possible alignment: $\mathbf{z} = \{(i_1, j_1), \ldots, (i_u, j_u)\}$:

```
x:          ...ATQRNDCAMFKRRPPSP--EGEHIL...
               |||  |||||   ||||  |||||||
y:          ...AAQ--DCEMF---PPAPREEGDHIL...
```

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

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

Null Distribution: $\mathbf{x}$, $\mathbf{y}$ iid $\sim \mu$

Possible alignment: $\mathbf{z} = \{(i_1, j_1), \ldots, (i_u, j_u)\}$:

```
x:      ...ATQRNDCAMFKRRPPSP--EGEHIL...
           |||  |||||   ||||  ||||||
y:      ...AAQ--DCEMF---PPAPREEGDHIL...
```

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

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

Null Distribution: $\mathbf{x}$, $\mathbf{y}$ iid $\sim \mu$

Two Questions:

How does $H_n(\mathbf{x}, \mathbf{y})$ grow with n?

What is $\mathbb{P}_0(H_n(\mathbf{x}, \mathbf{y}) > b)$ for large b?

Basic Result: The Phase Transition Phenomenon

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

By the theory of subadditive sequences,

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

Basic Result: The Phase Transition Phenomenon

Let $G_n(\mathbf{x},\mathbf{y})$ be the maximum alignment score of $\mathbf{x}$ and $\mathbf{y}$, *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

$$\alpha > 0 \quad \Rightarrow \quad \mathbb{P}(\lim_{n\to\infty} \frac{H_n}{n} = \alpha) \to 1$$

$$\alpha < 0 \quad \Rightarrow \quad \exists\, b \;\; s.t.\; \forall\, \epsilon > 0, \quad \mathbb{P}((1-\epsilon)b < \frac{H_n}{log(n)} < (2+\epsilon)b) \to 1$$

Brief Literature Review

Brief Literature Review

1. Gaps NOT Allowed:

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

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

$H_n(\mathbf{x},\mathbf{y})$ 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:

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

   $H_n(\mathbf{x}, \mathbf{y})$ 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:

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

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

2. Gaps Allowed: No complete theory.

   - Altschul and Gish (1996) *Methods in Enzymology*

Brief Literature Review

1. Gaps NOT Allowed:

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

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

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

2. Gaps Allowed: No complete theory.

   - Altschul and Gish (1996) *Methods in Enzymology*
   - 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:

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

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

2. Gaps Allowed: No complete theory.

   - Altschul and Gish (1996) *Methods in Enzymology*
   - 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:

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

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

2. Gaps Allowed: No complete theory.

   - Altschul and Gish (1996) *Methods in Enzymology*
   - 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:

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

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

2. Gaps Allowed: No complete theory.

   - Altschul and Gish (1996) *Methods in Enzymology*
   - Mott and Tribe (1999) *Journal of Computational Biology*
   - Siegmund and Yakir (2000) *Annals of Statistics*
   - Grossman and Yakir (2004) *Bernoulli*
   - Chan (2003) *Bernoulli*, (2005) Annals of Appl. Prob.

A Theorem from Chan (2003)

Let (K, Δ, δ) be chosen such that the convex function

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

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

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

A Theorem from Chan (2003)

Let (K, Δ, δ) be chosen such that the convex function

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

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

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

Works for $K =$Blosum62, $\Delta = 18$, $\delta = 1$.

Grossman and Yakir (2005)

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

$$f(x,y) = e^{\tilde{\theta}K(x,y)-s(\tilde{\theta})}\mu(x)\mu(y), \text{ where } s(\tilde{\theta}) = log\left(1 + 2\sum_{k\geq 1} e^{-\theta g(k)}\right).$$

   and (G_l^x, G_l^y), the gap at position l, from

$$\mathbb{P}((G_l^x, G_l^y) = (k,0)) = \mathbb{P}((G_l^x, G_l^y) = (0,k)) = e^{-\tilde{\theta}g(k)-s(\tilde{\theta})}$$

   Let $i_{l+1} = i_l + G_l^x$, $j_{l+1} = j_l + G_l^y$.

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

$$\mathbb{P}(H_n(\mathbf{x},\mathbf{y}) > b) = \mathbb{E}_Q\left[\frac{dP}{dQ}; H_n(\mathbf{x},\mathbf{y}) > b\right] \leq \mathbb{E}_Q\left[\frac{dP}{dQ_{z^*}}; H_n(\mathbf{x},\mathbf{y}) > b\right] \leq n^2 e^{-\tilde{\theta}b}$$

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: $\mathcal{A} = \{0, 1\}, \;\; K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

```
x:          ...1 1 0 1 1...

y:          ...1 1 1 0 0 0 1 1...
```

An optimal alignment is heavily constrained around gaps...

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

```
x:          ...1 1 0 1 1...

y:          ...1 1 1 0 0 0 1 1...


x:          ...1 1 1 1 1...

y:          ...1 1 1 0 1 1 1 1 ...
```

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: ϕ_L

We can apply the move r times: $\phi_L^r = \underbrace{\phi * \phi * \ldots \phi}_{r}$

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: ϕ_L

We can apply the move r times: $\phi_L^r = \underbrace{\phi * \phi * \ldots \phi}_{r}$

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

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

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

Theorem 1

Let (K, Δ, δ) be chosen such that the convex function

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

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

$$\mathbb{P}(H_n(\mathbf{x}, \mathbf{y}) \geq b) \leq n^2 B(\tilde{\theta}) e^{-\tilde{\theta} b}.$$

Sketch of Proof (1): Let

$$q(u,v,1) = q(u,v,0) = \sum_{\mathbf{x}\in\mathcal{A}^{u+v},\mathbf{y}\in\mathcal{A}^{u}} e^{\tilde{\theta}(K(\mathbf{x},\mathbf{y})-\Delta-\delta v)} I_L(\mathbf{x},\mathbf{y})\mu(\mathbf{x})\mu(\mathbf{y}),$$

then q is a probability measure on the set of all possible section types. Also, for $\mathbf{x} \in \mathcal{A}^{u+v},\ \mathbf{y} \in \mathcal{A}^{u}$, let

$$q_{u,v,0}(\mathbf{x},\mathbf{y}) = q_{u,v,1}(\mathbf{x},\mathbf{y}) = \frac{e^{\tilde{\theta}(K(\mathbf{x},\mathbf{y})-\Delta-\delta v)} I_L(\mathbf{x},\mathbf{y})\mu(\mathbf{x})\mu(\mathbf{y})}{q(u,v,1)},$$

then $q_{u,v,\cdot}$ is a probability measure on the sequences of section type $(u,v,\cdot)$.

Construct Q on $\mathcal{A}^m \times \mathcal{A}^n$ 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 $(\mathbf{x},\mathbf{y})$ generated by alignment z. Let $Q = \sum_{z\in\mathcal{Z}} Q_z$.

<ins>Sketch of Proof (2):</ins> Let z have sections $\{C_k : 1 \leq k \leq l\}$. By construction of Q_z, we have:

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

Sketch of Proof (2): Let z have sections $\{C_k : 1 \le k \le l\}$. By construction of Q_z, we have:

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

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

$$\begin{aligned}
\frac{dQ}{dP}(\mathbf{x},\mathbf{y}) &> \frac{\sum_{z \in \Phi(z^*)} dQ_z}{dP}(\mathbf{x},\mathbf{y}) \\
&> \left[\prod_{k=1}^{l} N_L(C_k)\right] \frac{dQ_{z^*}}{dP}(\mathbf{x},\mathbf{y}) \\
&= b(\tilde{\theta}) e^{\tilde{\theta} S_{z^*}(\mathbf{x},\mathbf{y})}/n^2 \\
&> b(\tilde{\theta}) e^{\tilde{\theta} b}/n^2
\end{aligned}$$

Therefore,

$$P(A) = \mathbb{E}_Q(\frac{dP}{dQ}, A) < \frac{1}{n^2} b^{-1}(\tilde{\theta}) e^{-\tilde{\theta} b}$$

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, ϕ_R.

For two adjacent sections C_1 and C_2

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

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

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 $T : \mathcal{Z}^+ \times \mathcal{Z}^+ \times \{0,1\} \rightarrow \mathcal{Z}^+$ be any 1-1, onto map. Let $\mathcal{M}(\theta) \doteq \mathcal{M}(\theta, K, \Delta, \delta)$ be the matrix with elements

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

If (K, Δ, δ) are chosen such that there exists a value of θ 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

$$\mathbb{P}(H_n(\mathbf{x},\mathbf{y}) \geq b) \leq n^2 B(\tilde{\theta}) e^{-\tilde{\theta} b}.$$

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 $D = diag(v^R)$.
2. Let $\pi' = [v_1^R v_1^L, v_2^R v_2^L, \ldots]$, 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 $\mathcal{M}(\tilde{\theta})$.

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(. . .) for all section types.

Thus we cap the number of allowable transforms to a maximum of κ. 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(. . .) for all section types.

Thus we cap the number of allowable transforms to a maximum of κ. 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.

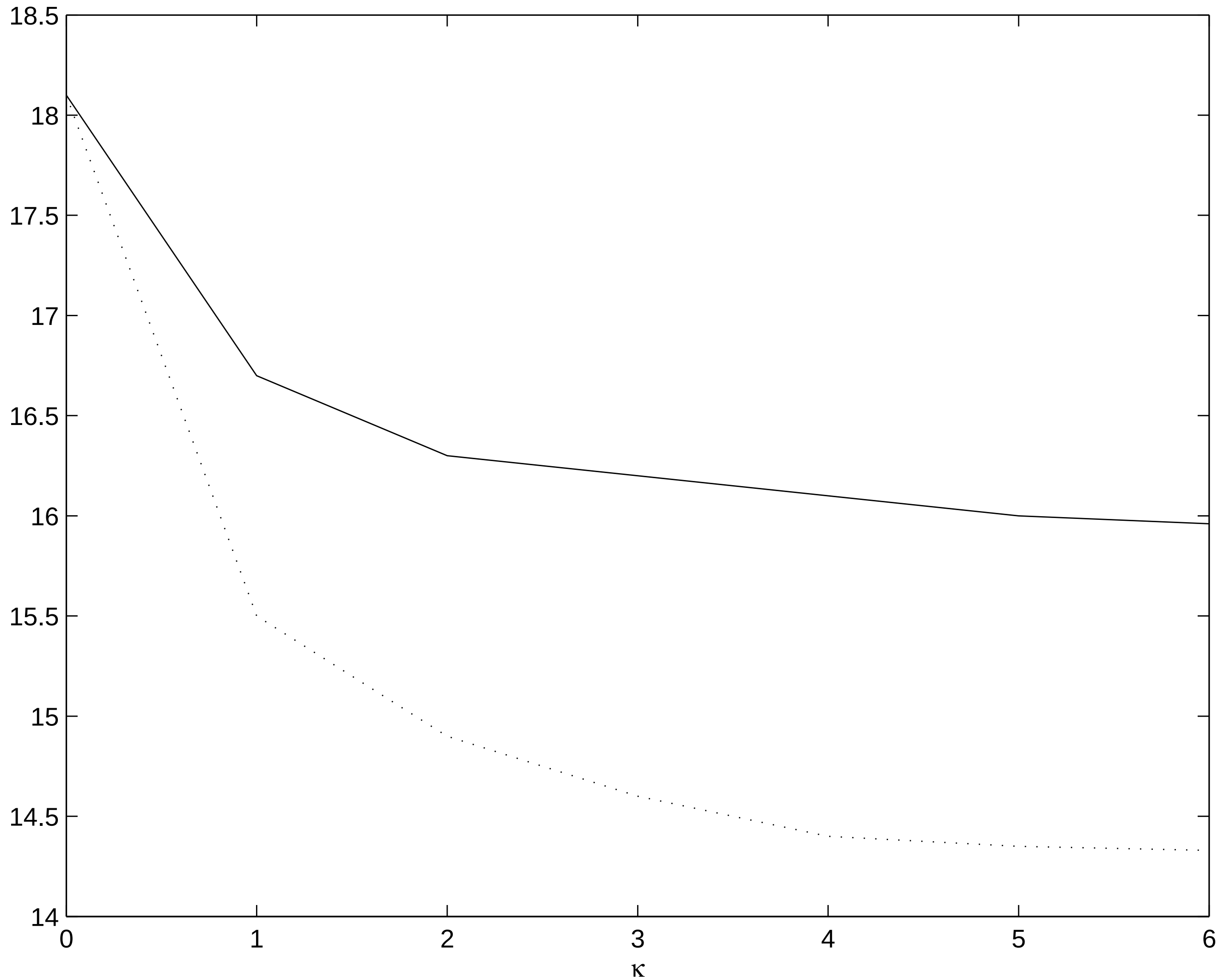


Figure 1. For $\delta = 1$ and increasing values of κ,the minimum value of Δ that can be proven to be in the logarithmic region. Dashed line is for non-Markov result, solid line is for Markov sections result.

How well can we possibly do using local optimality?

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

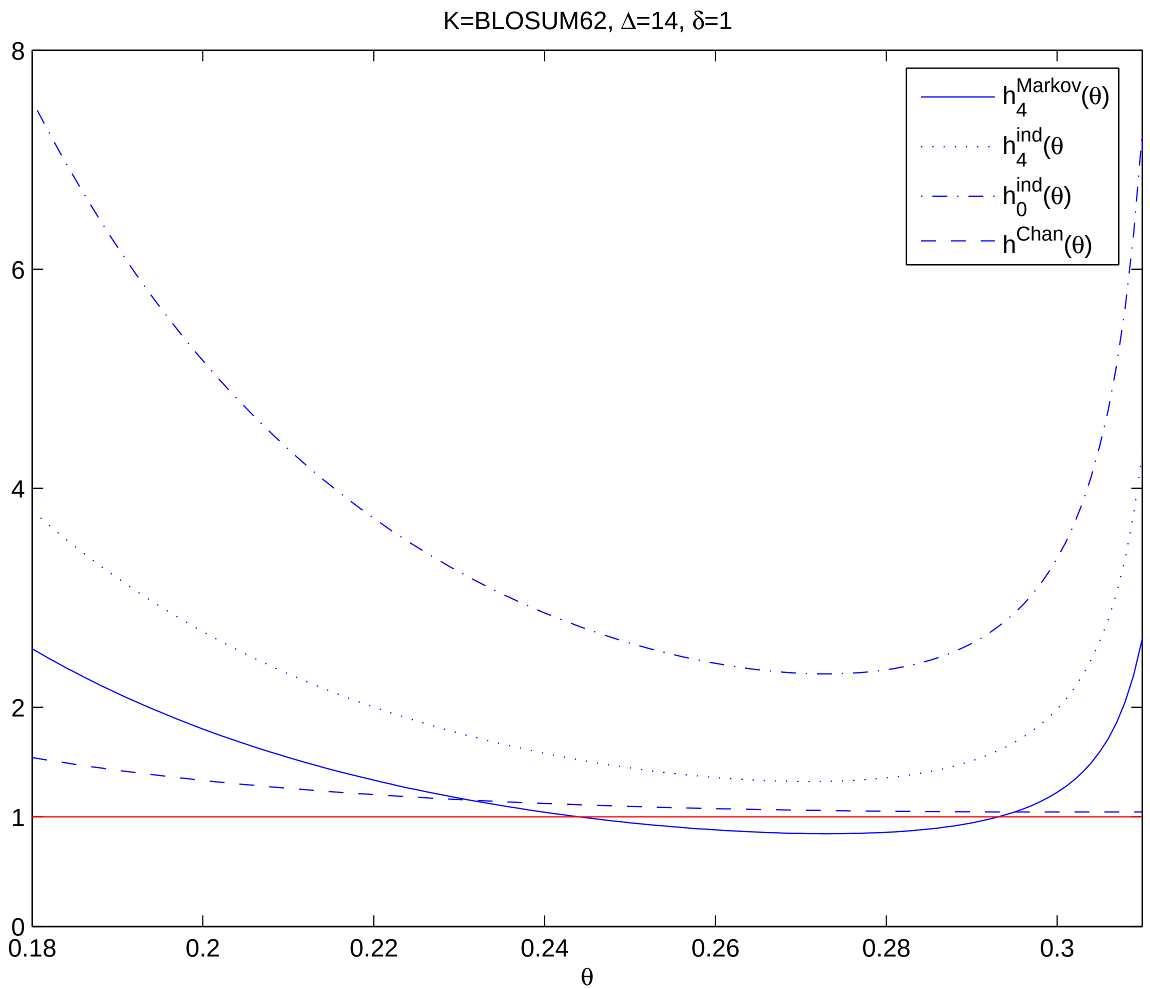
$$E_z[I^\kappa(z)] = E_z\left[\frac{\max_\phi e^{\theta K(\phi(z))}}{\sum_\phi e^{\theta K(\phi(z))}}\right]$$

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 $K =$ BLOSUM62, $\Delta = 15, \delta = 1.$



δ	Chan 2003	Independent Sections	Markov Sections	Altscul and Gish (1996)
1	18.1	16.1	14.4	≈ 8
2	15.0	13.0	11.3	≈ 6
3	13.0	10.9	9.2	≈ 5

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.

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!