A Practical Approach to Significance Assessment in Alignments with Gaps

Ralf Bundschuh
Department of Physics, The Ohio State University

Collaborators:
N. Chia, Department of Physics, Ohio State University
T. Hwa, R. Olsen, Physics Department, UC San Diego
S. Altschul, Y.K. Yu, National Center for Biotechnology Information, NIH

Outline:
- Biological sequences and what to do with them
- Sequence alignment
- Statistical significance
- Gapless alignment
- Alignment with gaps (analytical solution)
- Alignment with gaps (fast numerics)
- Conclusions

Funding: NSF, DAAD, Beckman foundation
The data: a piece of human chromosome 21

Sequence of 3 billion A,C,G, and T’s

The problem: What does it mean?
Biological sequences and what to do with them II

- How do we find out what a sequence does?

- **First step**: Determine which part codes for proteins and determine protein sequence \(\rightarrow\) solved, e.g., myoglobin: `MGLSDGEWQLVLHVWAKVEDVAGHGQDILIRLF`...

- **Second step**: compare sequence to sequences of genes (of other organisms) with known function \(\Rightarrow\) Sequence alignment algorithms

\[\Sigma=20\]

Calculate score \(\Sigma\) for every pair of sequences.

global: *Needleman and Wunsch, 1970*, local: *Smith and Waterman, 1981*

- Sequences similar enough \(\Rightarrow\) putative functional similarity

- Putative function to be confirmed by experiment

- Everybody does it: paper presenting most widely used program **BLAST** is most cited paper in all of science written in the 90’s

*Altschul et al., 1990*
**Sequence alignment I**

- How do sequence alignment algorithms work?
- Want to find sequence similarities, e.g., \text{AGMKCYDHP SARQAW KDAGVMK YEHP SQRW}
- Need way to compare individual letters \rightarrow scoring matrix \( s_{a,b} \)
  - Has to represent similarities between letters
  - DNA-DNA comparison:
    \[
    s_{a,b} = \begin{cases}
    1 & \text{for } a = b \\
    -\mu & \text{for } a \neq b
    \end{cases}
    \]
  - Protein-protein: BLOSUM or PAM matrices
    - Many parameters
  - Combine entries from the scoring matrix to an alignment score \( \Sigma \)
Sequence alignment II

- Simplest algorithm: gapless alignment

- Given pair of sequences $a_1 \ldots a_N$ and $b_1 \ldots b_N$

- Assign score to pair of substrings $a_{i-\ell+1} \ldots a_i$ and $b_{j-\ell+1} \ldots b_j$

\[
S[i, j, \ell] \equiv \sum_{k=0}^{\ell-1} s_{a_{i-k}, b_{j-k}} \\
S(11, 13, 6) = 10 + 3 + 6 + 6 + 2 + 0 = 27
\]

- Optimal alignment has score $\Sigma \equiv \max_{i, j, \ell} S[i, j, \ell]$

- Clever way to find optimal alignment $S_{i,j} \equiv \max_{\ell} S[i, j, \ell]$

\[
\Rightarrow \quad S_{i,j} = \max\{S_{i-1,j-1} + s_{a_i,b_j}, 0\} \\
\Sigma = \max_{i,j} S_{i,j}
\]
Problem: Algorithms assign score $\Sigma$ to every sequence pair, even random ones

Which $\Sigma$ implies biological relationship?

Statistical answer: find score distribution $P(\Sigma)$ if algorithm is applied to random sequences

\[ P(\Sigma) = \text{probability distribution of scores} \]

\[ \Rightarrow \text{can assign probability } p(\sigma) = \Pr\{\Sigma \geq \sigma\} = \int_{\sigma}^{\infty} P(\Sigma)d\Sigma \text{ to each score} \]

Choose highest score across a database of size $N$

- $p > 1/N \Rightarrow$ not similar enough
- $p < 1/N \Rightarrow$ biologically relevant!
• Need score distribution $P(\Sigma)$ for random sequences
• Obtain distribution by numerical simulation (shuffling method)

- Generate many pairs of random sequences with correct letter composition
- Align each pair
- Take histogram of alignment scores

• Very time consuming (hours)
Precompute distribution

But: distribution depends on

- Scoring parameters, e.g. $s_{a,b}$
- Sequence ensemble, e.g. $p_a$

Pre-computation only possible for some fixed set of scoring systems at overall amino acid/base pair frequencies

Problems:

- unusual amino acid composition
- Iterative schemes (PSI-BLAST)

- Position dependent scoring parameters

$\Rightarrow$ Need fast numerical way or analytical theory to get score distribution
Analytical theory for gapless alignment: Gumbel or extreme value distribution


\[ P(\Sigma) = \kappa \lambda \exp[-\lambda \Sigma - \kappa e^{-\lambda \Sigma}] \]

- universal shape
- Dependence on scoring system and letter frequencies in parameters \( \kappa, \lambda \)
- \( \kappa, \lambda \) known: \[ \langle e^{\lambda s} \rangle \equiv \sum_{a,b} p_a p_b e^{\lambda s_{a,b}} = 1 \]
- Corrections due to finite sequence length analytically known
- Makes gapless alignment powerful tool: BLAST

Altschul et al., J. Mol. Biol. 1990
Gapless Alignment I

- Basics of Karlin-Altschul theory:
  - random sequences:
    \[ S_{i,i} \rightarrow S(t) \]
    \[ s_{a,b} \rightarrow s(t) = \begin{cases} 
    1 & \text{with probability } p = 1/c \\
    -\mu & \text{with probability } 1 - p 
    \end{cases} \]
    \[ S(t + 1) = \max\{S(t) + s(t), 0\} \]
    \[ \Sigma = \max_t S(t) \]
  - two phases
  - \[ \langle s \rangle > 0 \]
    - Free random walk \[ S(t + 1) = S(t) + s(t) \]
    - \[ \langle \Sigma \rangle \sim N \]
    - linear alignment phase
    - Gaussian distribution of \( \Sigma \)
    - no detectability of local similarities
\[ S(t + 1) = \max\{S(t) + s(t), 0\} \]
\[ \Sigma = \max_t S(t) \]

- \(\langle s \rangle < 0\)
- negative drift → cutoff at zero important
- statistically independent islands
- Island peak score: \(\sigma_k \Rightarrow \Sigma = \max_k \sigma_k\)
- Assume exponential distribution of island peak scores
  \[ \Pr\{\sigma_k > \sigma\} = e^{-\lambda \sigma} \]
- for \(\kappa = N/\langle \ell \rangle\) such islands:
  \[ \Pr\{\Sigma < S\} = \Pr\{\sigma < S\}^\kappa = (1 - e^{-\lambda S})^\kappa \approx \exp[-\kappa e^{-\lambda S}] \]
- \(\langle \Sigma \rangle \sim \log N \rightarrow \text{logarithmic phase} \)
Need island peak score distribution

- Probability to reach peak score $\sigma$ within a given length $\ell$

$$p(\sigma | \ell) = p\left(\sum_{t=1}^{\ell} s(t)\right) = \frac{1}{2\pi i} \int e^{-\lambda \sigma} \langle e^{\lambda} \sum_{t=1}^{\ell} s(t) \rangle d\lambda = \frac{1}{2\pi i} \int e^{-\lambda \sigma} \langle e^{\lambda} \rangle^\ell d\lambda \equiv \frac{1}{2\pi i} \int e^{-\lambda \sigma} \rho(\lambda) d\lambda$$

- Probability maximal at some typical length $\ell_0(\sigma) \Rightarrow p(\sigma) \approx p(\sigma | \ell_0(\sigma))$

- Assume typical island slope: $\sigma = \alpha \ell_0(\sigma)$

$$p(\sigma) \sim p(\sigma | \ell = \alpha^{-1} \sigma) \sim \int \exp\{-[\lambda - \alpha^{-1} \log \rho(\lambda)] \sigma\} d\lambda$$

- Main contribution for large $\sigma$

$$p(\sigma) \sim \exp\{-[\lambda^*(\alpha) - \alpha^{-1} \log \rho(\lambda^*(\alpha))] \sigma\}$$

with $\lambda^*(\alpha)$ given by $\frac{d}{d\lambda} \big|_{\lambda = \lambda^*(\alpha)} [\lambda - \alpha^{-1} \log \rho(\lambda)] = 0$ \Rightarrow \alpha = \frac{\rho'(\lambda^*(\alpha))}{\rho(\lambda^*(\alpha))}$

- Find correct slope $\alpha^*$ by maximizing $p(\sigma)$ with respect to $\alpha$

$$0 = \log \rho(\lambda^*(\alpha^*)) = \log \langle e^{\lambda^*(\alpha^*)} \rangle \Rightarrow p(\sigma) \sim e^{-\lambda^*(\alpha^*) \sigma}$$
Problem: during evolution pieces of sequences are \textit{inserted} and \textit{deleted} ⇒ alignment algorithm has to \textit{compensate} for this ⇒ do gapped alignments to find \textit{weak} similarities

- First: gapped \textit{global} alignment

Alignment = way of inserting \textit{gaps} in sequences

\begin{align*}
--AG-MKCYDHPSARQAW \\
KDAGVMK-YEHPSS--QRW
\end{align*}

- Score gaps by $-\delta$

- Score $S[\mathcal{A}]$ for each alignment $\mathcal{A}$

$$S[\mathcal{A}] = \sum_{(a,b) \in \mathcal{A}} s_{a,b} - \delta N_g$$

- In practice: affine gap cost $\delta + \epsilon k$

- Find alignment with \textit{highest score}

$$\Sigma = \max_{\mathcal{A}} S[\mathcal{A}]$$

- Exponentially many alignments for each pair of sequences!
• Representation on an alignment grid:

\begin{align*}
&\text{AG-MKCYDHPQARQA}W \\
&\text{KDAGVMK-YEHPSQRW}
\end{align*}

• Dynamic programming: $O(N^2)$ algorithm

Needleman and Wunsch, 1970

\[ h(r, t + 1) = \max\{h(r \pm 1, t) - \delta, h(r, t - 1) + s(r, t)\} \]

• $\Sigma = h(0, 2N)$
**Sequence alignment V**

- Often only pieces of two sequences related ⇒ local alignment with gaps
  - Look at all pairs of subsequences $a_i \ldots a_j$ and $b_k \ldots b_l$ of the original sequences
  - Calculate gapped global alignment score $\Sigma(i, j, k, l)$ for each pair using Needleman-Wunsch algorithm
  - Total alignment score is best of these $\Sigma = \max_{i,j,k,l} \Sigma(i, j, k, l)$

- $O(N^6)$ algorithm?

- Again dynamic programming solution

  Smith and Waterman, 1981

$$
S_{j,l} \equiv \max_{i,k} \Sigma(i, j, k, l)
$$

$$
S_{j,l} = \max\{S_{j-1,l} - \delta, S_{j,l-1} - \delta, S_{j-1,l-1} + s_{a_j,b_k}, 0\}
$$

$$
\Sigma = \max_{j,k} S_{j,k}
$$

- Practically used sequence comparison programs (BLAST, FASTA) are approximations of this algorithm
Problem: Significance assessment less clear for alignment with gaps

Again phase transition:

- if $S$ tends to grow: global alignment, $\langle \Sigma \rangle \sim N$
- if $S$ does not grow: $\langle \Sigma \rangle \sim \log N$
- due to optimization $S$ can grow even for $\langle s \rangle < 0$
- where is phase transition?

Shape of distribution {\textbf{numerically verified}} to be still Gumbel

What are the values of the non-universal parameters $\lambda$ and $\kappa$?

Only very \textbf{limited theories and heuristics}

$S(r, t + 1) = \max \{S(r \pm 1, t) - \delta, S(r, t - 1) + s(r, t), 0\}$

$\Sigma = \max_{r, t} S(r, t)$

- Mark local path (---) for every lattice point with $S(r, t) > 0$
- Combine lattice sites which go back to same initiation event (○) to islands
- Islands disconnected from each other
- Island peak (●) score $\sigma_k$

$p(\sigma_k)^{\Sigma=\max_k\{\sigma_k\}} \rightarrow p(\Sigma) \Rightarrow \text{Gumbel distribution}$
• $\Sigma = \max_k \{\sigma_k\} \rightarrow \lambda, \kappa$ of distribution of $\Sigma$ given by island peak score distribution.

• Practical application: Measure island peak score distribution $\rightarrow$ deduce $\lambda, \kappa$

Olsen, RB, and Hwa, ISMB 1999; Altschul, RB, Olsen, and Hwa, NAR 2001
\( S(r, t + 1) = \max\{S(r, t - 1) + s(r, t), S(r \pm 1) - \delta, 0\} \)

- Need to know island peak score distribution
- Island of length \( \ell \) looks like global alignment
  \( h(r, t + 1) = \max\{h(r, t - 1) + s(r, t), h(r \pm 1, t) - \delta\} \)

- Island peak score \( \sigma = h(0, \ell) \)
- Define \( \rho(\lambda) \equiv \langle e^{\lambda h(0, \ell)} \rangle^{1/\ell} \Rightarrow \text{Gumbel } \lambda \text{ given by} \)
  \[ \rho(\lambda) = 1 \iff 0 = \Phi(\lambda; \Omega) \equiv \log \rho(\lambda) = \lim_{\ell \to \infty} \log \langle e^{\lambda h(0, \ell)} \rangle \]
  with \( \Omega = (s_a, \delta, p_a) \)
Alignment with gaps (analytical) I

- Simplest model:
  \[ s_{a_i, b_j} = \begin{cases} 
  1 & a_i = b_j \\
  0 & a_i \neq b_j 
\end{cases} \]

- longest common subsequence

Approximate by uncorrelated disorder

[Diagram of a rectangular lattice]

- Rectangular lattice of width 2W
- periodic boundary conditions
- at the end, \( W \to \infty \)
- \( h(r, t + 1) = \max \{ h(r, t - 1) + \eta(r, t), h(r \pm 1, t) \} \)

- Calculate \( \Phi(\lambda; \Omega) = \lim_{\ell \to \infty} \frac{1}{\ell} \log \langle e^{\lambda h(0, \ell)} \rangle \)
Alignment with gaps (analytical) II

\[ h(r, t + 1) = \max\{h(r, t - 1) + \eta(r, t), h(r \pm 1, t)\} \]

- **Microscopic dynamics**

- **Bond variables** \( n(r, t) \)

- **Dynamics of** \( h \) \( \rightarrow \) **dynamics of** \( n \)
  - \( n(r, t) \) independent of absolute score \( h \)
  - \( n(r, t) \in \{0, 1\} \)
  - In one time step, pairs of neighboring \( n(r, t) \) get transformed into pairs of neighboring \( n(r, t) \)
  - Transformation of pair \( (n'_1, n'_2) \) into pair \( (n_1, n_2) \) described by transfer matrix \( T_0 \)

- \( n(r, t) \) perform a Markov process on the \( 2^{2W} \)-dimensional state space \( \{0, 1\}^{2W} \)
Set up generalized transfer matrix $T_W(\lambda)$ for fixed $W$ (factor $e^\lambda$ for each transition that increases $h$ by one)

Solve for largest eigenvalue $\rho_W(\lambda)$ for small $W$ using computer algebra $\Rightarrow \Phi_W(\lambda; \Omega) = \log \rho_W(\lambda)$

Generalize for arbitrary $W$

Take limit

$$\Phi(\lambda; \Omega) = \lim_{W \to \infty} \Phi_W(\lambda; \Omega) = \log \frac{\exp[-\frac{\lambda}{2}] + \sqrt{p}}{1 + \sqrt{p} \exp[-\frac{\lambda}{2}]}$$

For general $\mu = 2\delta \neq 0$

$$\Phi(\lambda; \Omega) = \log \left( \frac{1 + \sqrt{p} \exp[\frac{\lambda}{2}(1 + \mu)]}{1 + \sqrt{p} \exp[-\frac{\lambda}{2}(1 + \mu)]} \right) - \frac{\lambda}{2\mu}$$

Gumbel $\lambda$ given by

$$\log \left( \frac{1 + \sqrt{p} \exp[\frac{\lambda}{2}(1 + \mu)]}{1 + \sqrt{p} \exp[-\frac{\lambda}{2}(1 + \mu)]} \right) - \frac{\lambda}{2\mu} = 0$$
• Compare calculated with directly simulated $\lambda$

• $p = 1/4$ (DNA alphabet)

- Perfect agreement for uncorrelated disorder
- Disorder correlations have only weak influence
- Difference between correlated and uncorrelated case smaller as $p \to 0$ (larger alphabet)
• More general case: \(2\delta + 1 = m(\mu + 1)\) for integer \(m\):

• Same geometry: bond variables \(n(r, t)\)

• Dynamics of \(h \rightarrow\) dynamics of \(n\)
  – \(n(r, t)\) independent of absolute score \(h\)
  – \(n(r, t) \in \{0, 1, 2, \ldots, m\}\)
  – In one time step, pairs of neighboring \(n(r, t)\) get transformed into pairs of neighboring \(n(r, t)\)
    – Transformation of pair \((n'_1, n'_2)\) into pair \((n_1, n_2)\) described by transfer matrix \(T_0\)

• \(n(r, t)\) perform a Markov process on the \((m + 1)^{2W}\)-dimensional state space \(\{0, 1, 2, \ldots, m\}^{2W}\)
• Set up generalized transfer matrix $T_W(\lambda)$ for fixed $W$ (factor $e^\lambda$ for each transition that increases $h$ by one)

• Cannot solve for largest eigenvalue $\rho_W(\lambda)$ with computer algebra any more

• Can precalculate shape of $T_W(\lambda)$ for different $m$ with $\mu$ and $p$ as parameters using computer algebra

• Can calculate $\rho_W(\lambda)$ (and thus $\Phi_W(\lambda; \Omega)$) numerically very efficiently using high performance numerics package ARPACK (ARNOLDI method)

Lehoucq et al., SIAM 1997

• Still need to take limit $W \to \infty$

• Use concept of universality
Alignement with gaps (fast numerics) III

- **Universality:** all systems in one universality class share certain features

- Here: Kardar-Parisi-Zhang universality class

- Contains:
  - Models of surface growth
  - Polymer models
  - Traffic models
  - Asymmetric Exclusion Process (ASEP)
  - Sequence alignment with gaps

- Result of Derrida and Lebowitz (PRL 1998) for ASEP:

\[
\Phi(\lambda; \Omega) = \Phi_W(\lambda; \Omega) - \frac{a_\Omega G(W^{1/2}b_\Omega \lambda)}{W^{3/2}}
\]

- \( G \) is universal scaling function, i.e., is the same for all systems

- \( a_\Omega \) and \( b_\Omega \) are non-universal scaling factors, i.e., they depend on the system and on the parameters \( \Omega = (\mu, \delta, p) \), but not on \( W \)
\[ \Phi(\lambda; \Omega) = \Phi_W(\lambda; \Omega) - \frac{a_\Omega G(W^{1/2} b_\Omega \lambda)}{W^{3/2}} \]  

\[ (*) \]

- Need to know \( a_\Omega \) and \( b_\Omega \)
- Fit to

\[ \Phi_W(\lambda; \Omega) - \Phi_{W-1}(\lambda; \Omega) = \frac{a_\Omega G(W^{1/2} b_\Omega \lambda)}{W^{3/2}} - \frac{a_\Omega G'((W - 1)^{1/2} b_\Omega \lambda)}{(W - 1)^{3/2}} \]  

\[ (**) \]

- Full procedure:
  - Precompute shape of generalized transfer matrix for different \( m \) using computer algebra
  - For given \( \Omega = (\mu, \delta, p) \) numerically solve for largest eigenvalue and obtain \( \Phi_W(\lambda; \Omega) \) and \( \Phi_{W-1}(\lambda; \Omega) \)
  - Obtain \( a_\Omega \) and \( b_\Omega \) by fitting to \( (**) \)
  - Obtain \( \Phi(\lambda; \Omega) \) from \( (*) \)
  - Solve for \( \Phi(\lambda; \Omega) = 0 \)
• Does it converge?

• Test on an $m = 1$ case where exact result is known

• Convergence already for very small $W$
• Performance:

<table>
<thead>
<tr>
<th>m</th>
<th>µ</th>
<th>δ</th>
<th>W</th>
<th>time (sec)</th>
<th>error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.2</td>
<td>1.1</td>
<td>5</td>
<td><strong>0.2</strong></td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>1.4</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>4.7</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>26.4</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>2.0</td>
<td>3</td>
<td><strong>0.1</strong></td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>2.2</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>26.5</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>2.35</td>
<td>4</td>
<td><strong>0.3</strong></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>108.2</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>2.9</td>
<td>3</td>
<td><strong>0.1</strong></td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td><strong>1.0</strong></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>39.0</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.7</td>
<td>3.75</td>
<td>2</td>
<td><strong>0.1</strong></td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0.2</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>5.3</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td>6</td>
<td>0.55</td>
<td>4.15</td>
<td>2</td>
<td><strong>0.1</strong></td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1.1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td><strong>39.5</strong></td>
<td>&lt; 0.1</td>
</tr>
</tbody>
</table>

\[2\delta = m(1 + \mu) - 1\]

• precise and very fast
Conclusions and outlook

- Sequence alignment is **standard tool** in molecular biology

- Sequence alignment algorithms rely on **dynamic programming methods**

- **Statistics** of sequence alignments are **important and poorly understood**

- Progress can be made through a **synthesis** of high performance numerics and methods from statistical physics

- **Future directions**
  - Other analytical solutions for $\lambda$
  - Affine gap costs
  - Protein scoring parameters
Importance Sampling

- Smart numerics

- Estimate $\langle e^{\lambda h(0, 2N)} \rangle$ by numerical sampling

- Instead of random sequence pairs (with $\langle h(0, 2N) \rangle < 0$) use correlated sequence pairs with $\langle h(0, 2N) \rangle > 0$ such that
  
  - $\langle e^{\lambda h(0, 2N)} \rangle_{\text{corr}}$ is not dominated by rare events
  
  - The difference between $\langle \cdot \rangle$ and $\langle \cdot \rangle_{\text{corr}}$ can be handled analytically

- Estimate $\langle e^{\lambda h(0, 2N)} \rangle$ for $N = 60$, $N = 80$, $N = 100$ and extrapolate to $N \to \infty$

Results:

<table>
<thead>
<tr>
<th>Scoring system</th>
<th>$\lambda_{\text{reference}}$</th>
<th>$\lambda_{\text{this algorithm}}$</th>
<th>time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM45/14/2</td>
<td>0.1961±0.0008</td>
<td>0.1978±0.0005</td>
<td>3:00</td>
</tr>
<tr>
<td>BLOSUM62/11/1</td>
<td>0.2670±0.0002</td>
<td>0.2669±0.0006</td>
<td>5:49</td>
</tr>
<tr>
<td>BLOSUM80/10/1</td>
<td>0.2993±0.0012</td>
<td>0.3004±0.0005</td>
<td>1:00</td>
</tr>
<tr>
<td>PAM70/10/1</td>
<td>0.2921±0.0013</td>
<td>0.2922±0.0003</td>
<td>0:26</td>
</tr>
<tr>
<td>PAM30/9/1</td>
<td>0.2963±0.0011</td>
<td>0.2954±0.0002</td>
<td>0:27</td>
</tr>
</tbody>
</table>
Hybrid Alignment I

- $\langle e^{\lambda h(0,2N)} \rangle = 1$ in general hard to fulfill

Instead of calculating complicated quantity for Smith-Waterman alignment change the algorithm

- Smith-Waterman

$$S(r,t+1) = \max \left\{ S(r+1,t) - \delta, S(r-1,t) - \delta, S(r,t-1) + s(r,t), 0 \right\}, \quad \Sigma = \max_{r,t} S(r,t)$$

- Replace by hybrid algorithm

$$Z(r,t+1) = Z(r+1,t)e^{-\lambda_{gl}\delta} + Z(r-1,t)e^{-\lambda_{gl}\delta}, \quad \Sigma = \max_{r,t} \log Z(r,t)$$

$$+ Z(r,t-1)(1 - 2e^{-\lambda_{gl}\delta})e^{\lambda_{gl}s(r,t)} + 1$$

- Similar to:

  Viterbi $\rightarrow$ probabilistic HMM (forward–backward)

  RNA minimal energy (Zuker) $\rightarrow$ partition function (Vienna)

- Guarantees $\langle e^{1 h(0,2N)} \rangle = 1 \implies \lambda = 1$ independent of scoring system
- Numerical test of *hybrid* statistics
- 5000 i.i.d. amino acid sequences of length $N$, PAM-120 scoring matrix, $11 + k$ gap cost

- Score histogram is of **Gumbel form**
- $\lambda = 1$ for large $N$
- Even sequence length dependence theoretically understood
- Works even for **position-specific** scoring systems
- E.g., protein family Hidden Markov Models from Pfam database

Bateman *et al.*, Nucleic Acids Res. 2000

- Score histogram is of **Gumbel form**
- $\lambda = 1$ within $\pm 10\%$ for all 2,216 models
How is the performance in terms of sensitivity?

Test algorithm on standard database: PDB90D-B

Use SCOP as “gold standard” (known relations between sequences)

- Do pairwise alignments of all sequences in database
- Vary \( p \)-value cutoff and measure

\[
\text{Coverage} = \frac{\text{Number of relations found}}{\text{Number of total relations}}, \quad \text{Errors per Query} = \frac{\text{Number of wrong relations}}{\text{Number of sequences}}
\]

- Ideal: high coverage at low errors per query

Hybrid alignment’s sensitivity is at least as good as that of other methods