Statistical Assessment of Sequence Alignments

Ralf Bundschuh
Department of Physics, The Ohio State University

Collaborators:
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
- Central conjecture
- Applications
- Conclusions

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

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 → solved, e.g., myoglobin: MGLSDGEWQLVLHVWAKVEADVAGHGQDILIRLF...
  - Second step: compare sequence to sequences of genes (of other organisms) with known function ⇒ Sequence alignment algorithms

![Diagram]

Calculate score Σ for every pair of sequences.


- Sequences similar enough ⇒ 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.,

\[
\text{AGMKCYDHP SARQAW} \quad \text{KDAGVMK YEHP SQRW}
\]
- Need way to compare individual letters → 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 \( \Sigma \)
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 $a_{i-l+1} \ldots a_i$ and $b_{j-l+1} \ldots b_j$

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

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

$$S[i,j,\ell] \equiv \sum_{k=0}^{\ell-1} s_{a_{i-k},b_{j-k}}$$

$$S(11,13,6)=10+3+6+6+2+0=27$$

$$\Rightarrow S_{i,j} = \max\{S_{i-1,j-1} + s_{a_i,b_j}, 0\}$$

$$\Sigma = \max_{i,j} S_{i,j}$$
**Statistical Significance I**

- **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.

⇒ can assign probability $p(\sigma) = \Pr\{\Sigma \geq \sigma\} = \int_\sigma^\infty P(\Sigma)d\Sigma$ 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)
Significance Assessment III

- 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

⇒ 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
**Sequence alignment III**

- Problem: during evolution pieces of sequences are inserted and deleted
  ⇒ alignment algorithm has to compensate for this ⇒ do gapped alignments to find weak similarities

- First: gapped global alignment

- Alignment = way of inserting gaps in sequences

  \[
  \text{AG-MKCYDHPSARQAW}
  \]

  \[
  \text{KDAGVMK-YEHPS--QRW}
  \]

- 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 highest score

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

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

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>M</th>
<th>K</th>
<th>C</th>
<th>Y</th>
<th>D</th>
<th>H</th>
<th>P</th>
<th>S</th>
<th>A</th>
<th>R</th>
<th>Q</th>
<th>A</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>K</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td>V</td>
<td>M</td>
<td>K</td>
<td>Y</td>
<td>E</td>
<td>H</td>
<td>P</td>
<td>S</td>
<td>Q</td>
<td>R</td>
<td>W</td>
</tr>
</tbody>
</table>
```

- 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

- 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

  \[ S_{j,l} \equiv \max_{i,k} \Sigma(i, j, k, l) \]
  \[ S_{j,l} = \max\{ S_{j-1,l} - \delta, S_{j,l-1} - \delta, S_{j-1,l-1} + s_{a_j,b_k}, 0 \} \]
  \[ \Sigma = \max_{j,k} S_{j,k} \]

- Practically used sequence comparison programs (BLAST, FASTA) are approximations of this algorithm
Significance Assessment V

- Problem: Significance assessment less clear for alignment with gaps
- Shape of distribution numerically verified to be still Gumbel
- What are the values of the non-universal parameters $\lambda$ and $\kappa$?
- Only very limited theories and heuristics
  

- Statistical physics problem:
  Find distribution of observable in disordered system

$\Rightarrow$ Use statistical physics methods
Central conjecture

- **Central conjecture:**

  The distribution of $\Sigma$ is of the Gumbel form with the Gumbel parameter $\lambda$ given by

  $$\langle e^{\lambda h(0,2N)} \rangle = 1$$

- **Properties:**
  
  - $h(0, 2N)$ free energy
    ⇒ $\langle e^{\lambda h(0,2N)} \rangle \approx \langle Z^\lambda \rangle$ “replicated” partition function
    ⇒ hard to calculate analytically
  
  - $\langle h(0, 2N) \rangle < 0$
    ⇒ typical sequence pairs contribute zero to $\langle e^{\lambda h(0,2N)} \rangle$
    ⇒ rare events dominate $\langle e^{\lambda h(0,2N)} \rangle$
    ⇒ numerically hard
  
  - Reduces to proven formula $\langle e^{\lambda s} \rangle = 1$ for gapless alignment
  
  - Not proven, but applications work
Applications I

- **Application I**: DNA-DNA comparison with $\mu = 2\delta$.
- Mapping possible to asymmetric exclusion process
  - Model of highway traffic
  - Exactly solvable model of non-equilibrium physics
- Can calculate $\langle e^{\lambda h(0,2N)} \rangle$
- Gumbel $\lambda$ given by
  \[
  \frac{1 + \sqrt{p} \exp\left[\frac{\lambda}{2}(1 + \mu)\right]}{1 + \sqrt{p} \exp\left[-\frac{\lambda}{2}(1 + \mu)\right]} \exp\left[-\frac{\lambda}{2}\mu\right] = 1 \quad (*)
  \]
- Compare calculated with simulated $\lambda$ ($p = 1/4$, DNA)
- Provides test case for heuristic approaches

$\Rightarrow$ BLAST $p$-value calculation changed since version 2.1.3

*Altschul, RB, Olsen, and Hwa, Nucleic Acids Res. 2001*
Applications II

- Application II: 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>
Application III:  $\langle e^{\lambda h(0,2N)} \rangle = 1$ 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

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

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

Similar to:

- Viterbi $\rightarrow$ probabilistic HMM (forward–backward)
- RNA minimal energy (Zuker) $\rightarrow$ partition function (Vienna)

Guarantees $\langle e^{\lambda h(0,2N)} \rangle = 1 \Rightarrow \lambda = 1$ 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
**Applications VI**

- How is the **performance** in terms of **sensitivity**?
- Test algorithm on **standard database**: PDB90D-B
  

- Use **SCOP** as “gold standard” (known relations between sequences)
  
  *Murzin et al., J. Mol. Biol. 1995*
  
  - 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

---

![Graph showing comparison of different alignment methods](image)
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**
- Statistical physics **methods** can be applied to characterize and **improve** sequence alignment

Future directions
- Other analytical solutions for $\lambda$
- Better importance sampling to estimate $\lambda$
- Incorporate hybrid algorithm into PSI-BLAST
- Statistics of other algorithms of computational biology (RNA folding, gene finding, clustering, . . .)