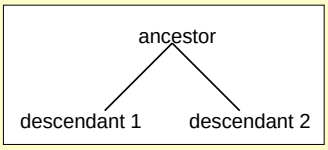
Distribution of Patterns in Pairwise Sequence Alignments

S. Wolfsheimer and G. Nuel

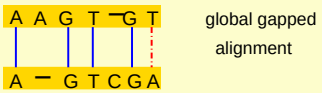
MAP5 – Mathématiques Appliquées á Paris 5, University Paris Descartes, France

Sequence Alignment [1]

Detection of homological relationships in DNA- or protein-sequences.
Applications: Search tools for molecular databases, e.g. BLAST



- Given: Pair of sequences $\mathbf{a} = a_1 ... a_M \in \Sigma^M$, $\mathbf{b} = b_1 ... b_N \in \Sigma^N$.
- amino acids: $\Sigma = \{A, C, D, E \ldots\}$
  nucleotides: $\Sigma = \{A, G, C, T\}$
- Alignment $\mathcal{A}$: set of pairings $(i_1, j_1), \ldots, (i_n, j_n)$ with $i_k < i_{k+1}$, $j_k < j_{k+1}$
- Gaps: insertions and deletions of subsequences



global gapped alignment

- Alignment score:

$$\mathsf{S}(\mathcal{A}; \mathbf{a}, \mathbf{b}) = \sum_{\text{pairs}} \sigma_{a_i, b_j} - \sum_{\text{gaps}} g(l_{\text{gap}})$$

- score matrix: $\sigma_{a,b} = \log \frac{P_{a,b}}{f_a f_b}$
  gap penalty: $g(l) = \alpha + \beta(l-1)$

Score based alignment: Similarity between $\mathbf{a}$ and $\mathbf{b}$ measured by optimal alignment score S_0.

- $S_0 = \max_{\mathcal{A}} \mathsf{S}(\mathcal{A}; \mathbf{a}, \mathbf{b})$
- $\mathcal{A}_0 = \text{argmax}_{\mathcal{A}} \mathsf{S}(\mathcal{A}; \mathbf{a}, \mathbf{b})$

Dynamic Programming

Global alignment:
Needleman-Wunsch algorithm

- Here: linear gap costs $g(l_{\text{gap}}) = \delta l_{\text{gap}}$
  Affine gap-costs straight forward.
- A_{ij} : optimal alignment of subproblem $a_1, \ldots a_i$, $b_1, \ldots b_j$

$$A_{i,j} = \max \begin{cases} A_{i-1,j-1} + \sigma_{a_i,b_j} \\ A_{i-1,j} - \delta \\ A_{i,j-1} - \delta \end{cases}$$

$$S_0 = A_{MN}$$

- $\mathcal{O}(M \cdot N)$ time complexity

Alignment biases [3]

Close to gaps often many competitive alignments decrease the accuracy of the score based alignment. Typical effects are:

- Gap wander: a gap shifted by a few positions.

- Gap attraction: two close gaps merge into a single gap.

- Gap annihilation: two gaps of opposite signature (insertion / deletion) cancel each other.

Probabilistic alignment: more quantitative description. Here: distributions of typical patterns that might lead to weakly reliable alignment segments.

Probabilistic alignment [2]

- $P_T(\mathcal{A})$: distribution of global alignments of $\mathbf{a}$, $\mathbf{b}$.
- Canonical ensemble of alignments $P_T(\mathcal{A}; \mathbf{a}, \mathbf{b}) = \frac{1}{Z_T} \exp\left[\frac{1}{T} S(\mathcal{A}; \mathbf{a}, \mathbf{b})\right]$ (Boltzmann distribution with "temperature" T = 1)
- Partition function: $Z_T = \sum_{\mathcal{A}} \exp\left[S(\mathcal{A})/T\right]$
- Partition function calculation: $D_{i,j} \to Z_{i,j}$, $\max \to \sum$, and $+ \to \times$
- Forward algorithm:
  $Z_{i,j}$: sum over all alignments of $a_1 \ldots a_i$ and $b_1 \ldots b_j$

$$Z_{i,j} = \underbrace{Z_{i-1,j-1}\, e^{\sigma_{a_i,b_j}/T}}_{\text{match/mismatch}} + \underbrace{(Z_{i-1,j} + Z_{i,j-1}) e^{-\delta/T}}_{\text{gap}}$$

$$Z_T = Z_{M,N}$$

- Backward algorithm:
  $Z'_{i,j}$: sum over all alignments of $a_{i+1} \ldots a_M$ and $b_{j+1} \ldots b_N$ given that $(i,j) \in \mathcal{A}$.
- Posterior probabilities:

$$p_{ij} := P[(i,j) \in \mathcal{A}] = \frac{1}{Z_T} Z_{i,j} Z'_{i,j}$$

$$P\begin{pmatrix} a_i \\ - \end{pmatrix} = 1 - \sum_j p_{ij}$$

Number of gaps

Distribution of the number of gaps (see [4] for a general HMM setup).

Insertion

$$\begin{matrix} a_{i-l} \cdots a_{i-1}\, a_i & & a_{i+1} \cdots a_{i+l'} \\ b_j & - - - - - & b_{j+l} \end{matrix}$$

Deletion

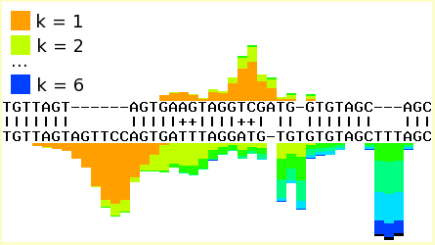
$$\begin{matrix} a_i & - - - - - & a_{i+1} \\ b_{j-l} \cdots b_{j-1}\, b_j & & b_{j+1} \cdots b_{j+l'} \end{matrix}$$

Insertions: auxiliary matrices $Z^{(I,k)}_{i,j}$ and $I^{(I,k)}_{i,j}$ for $k = 1, 2, \ldots$, sum over alignments, k insertions seen so far.

$I^{(I,k)}_{i,j}$: sum over $\begin{matrix} \cdots & \cdots & \cdots & a_i \\ \cdots & b_j & - & - \end{matrix}$

$$\begin{aligned} Z^{(I,k)}_{i,j} &= \left(I^{(I,k)}_{i-1,j-1} + Z^{(I,k)}_{i-1,j-1}\right) e^{\sigma_{a_i,b_j}/T} \\ &\quad + \left(Z^{(I,k)}_{i,j-1} + I^{(I,k)}_{i,j-1}\right) e^{-\delta/T} \\ I^{(I,k)}_{i,j} &= \left(Z^{(I,k-1)}_{i-1,j} + I^{(I,k)}_{i-1,j}\right) e^{-\delta/T} \\ P(k) &= \left(Z^{(I,k)}_{M,N} + I^{(I,k)}_{M,N}\right) / Z_{M,N} \end{aligned}$$

$$P\left(\begin{matrix} a_i \\ - \end{matrix}, \; k\text{th insertion}\right) = \sum_j \frac{I^{(I,k)}_{i,j} \cdot I'_{i,j+1}}{Z_{M,N}}$$



Gap Annihilation Pattern

Detecting candidates for gap annihilation. Search for the pattern:

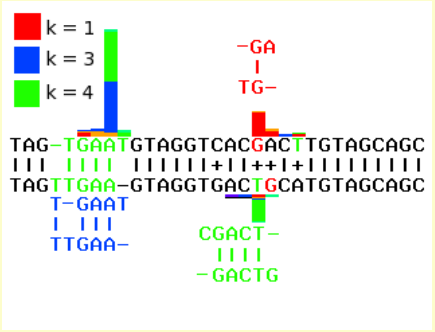
Insertion

$$\begin{matrix} \cdots & - & a_{i-k+1} \cdots a_i\, a_{i+1} \\ \cdots & b_{j-k}\, b_{j-k+1} \cdots b_j & - \end{matrix}$$

Deletion

$$\begin{matrix} \cdots & a_{i-k}\, a_{i-k+1} \cdots a_i & - \\ \cdots & - & \underbrace{b_{j-k+1} \cdots b_j}_{k}\; b_{j+1} \end{matrix}$$

Decode alternative alignment segments:



Bibliography

[1] *Durbin et.al.*, Biological sequence analysis (1998)
[2] *Miyazawa.*, Prot. Eng. **8** (1995)
[3] *Lunter et.al.* , Genome Research. **18** (2008)
[4] *Aston and Martin*, Ann. Appl. Stat.**1** (2007)