Multiple Sequence Alignment

Basics

Rolf Backofen

Lehrstuhl für Bioinformatik
Institut für Informatik
Albert-Ludwigs-Universität Freiburg

Course Bioinformatics I

built on February 6, 2019

Multiple Sequence Alignment (MSA)

Questions for today

1. How are multiple sequence alignments defined?
2. How are they scored?
3. How are they constructed?

Uses:

- helps to identify subsequences of functional importance
  e.g. protein domains, TF binding sites
- local or global similarity with biological meaning
- provides information on evolutionary history
  construction of phylogenetic tress
- technical: assembly of sequence reads after sequencing

MSA Definition

Definition (MSA)

Let $a^1 ... a^N$ be N sequences. A *multiple sequence alignment (MSA)* of $a^1 ... a^N$ is a matrix

$$\mathbf{A} = (A_{i,j}) \quad \begin{matrix} 1 \leq i \leq N & \leftarrow \text{sequences} \\ 1 \leq j \leq K & \leftarrow \text{columns of the MSA} \end{matrix}$$

with

1. $\forall i,j : A_{i,j} \in \Sigma \bigcup \{-\}$ (where Σ is the alphabet of the sequences)
2. $\forall i : (A_{i,1} ... A_{i,k}) \mid_\Sigma = a^i$ (sequence i in row i)
3. $\forall j : \text{not}(\forall i : A_{i,j} = -)$ (no columns with only gaps)

- **Example:**

$$\mathbf{A} = \begin{pmatrix} \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{C} \\ \mathtt{A} & \mathtt{A} & - & \mathtt{C} \\ - & \mathtt{A} & \mathtt{G} & - \end{pmatrix} \begin{matrix} 1 \\ \vdots \leftarrow i \\ N \end{matrix} \qquad \begin{matrix} 1 & \ldots & K \\ & j & \end{matrix}$$

$$a^1 = \mathtt{AAAC}$$
$$a^2 = \mathtt{AAC}$$
$$a^3 = \mathtt{AG}$$

Scoring of MSA: Definitions

Definition

Given an alignment **A** of N sequences with K columns. Furthermore, let $S_c : (\Sigma \bigcup \{-\})^N \rightarrow \mathbb{R}$ be a scoring function for columns. Then the score for the alignment **A** is defined by

$$S(\mathbf{A}) = \sum_{1 \leq j \leq K} S_c(A_{1j}, \ldots, A_{Nj})$$

- How to define $S_c(\ldots)$? $\Rightarrow$ recall PAM
  - **A** alignment of two sequences $\longrightarrow$ $S_c(x, y) = \log \frac{p_{xy}}{q_x q_y}$
  - **A** alignment of three sequences $\longrightarrow$ $S_c(x, y, z) = \log \frac{p_{xyz}}{q_x q_y q_z}$

$\Rightarrow$ problem with data, not enough alignments available

Sum-of-pairs (SP) Scoring

- Idea: define column score from pairwise scorings (e.g. using PAM), i.e.
  sum-of-pairs = SP (Carrillo&Lipman, 1988)
- Currently *the* standard method

Definition

$S_c(A_{1j}, \ldots, A_{Nj})$ is a *sum-of-pairs* score if $\exists\, s_p(A_{xj}, A_{yj})$ such that

$$S_c(A_{1j}, \ldots, A_{Nj}) = \sum_{1 \le k < l \le N} \underbrace{s_p(A_{kj}, A_{lj})}_{\text{using standard scoring}}$$

- **Basically:** score is the sum of all pairwise alignments

$$\begin{aligned} S\begin{pmatrix}\texttt{AAAC}\\ \texttt{AA-C}\\ \texttt{-AG-}\end{pmatrix} &= S_c\begin{pmatrix}\texttt{A}\\ \texttt{A}\\ \texttt{-}\end{pmatrix} + S_c\begin{pmatrix}\texttt{A}\\ \texttt{A}\\ \texttt{A}\end{pmatrix} + S_c\begin{pmatrix}\texttt{A}\\ \texttt{-}\\ \texttt{G}\end{pmatrix} + S_c\begin{pmatrix}\texttt{C}\\ \texttt{C}\\ \texttt{-}\end{pmatrix} \\ &= s_p\begin{pmatrix}\texttt{A}\\ \texttt{A}\end{pmatrix} + s_p\begin{pmatrix}\texttt{A}\\ \texttt{-}\end{pmatrix} + s_p\begin{pmatrix}\texttt{A}\\ \texttt{-}\end{pmatrix} \quad + s_p\begin{pmatrix}\texttt{A}\\ \texttt{A}\end{pmatrix} + s_p\begin{pmatrix}\texttt{A}\\ \texttt{A}\end{pmatrix} + s_p\begin{pmatrix}\texttt{A}\\ \texttt{A}\end{pmatrix} \quad + .. \\ &= S\begin{pmatrix}\texttt{AAAC}\\ \texttt{AA-C}\end{pmatrix} + S\begin{pmatrix}\texttt{AAAC}\\ \texttt{-AG-}\end{pmatrix} + S\begin{pmatrix}\texttt{AA-C}\\ \texttt{-AG-}\end{pmatrix} \end{aligned}$$

Problem with SP (Altschul, Carroll & Lipman, 1989)

- remember correct scoring extension from 2 to 3 sequences

$$S_c(x,y,z) = \log \frac{p_{xyz}}{q_x q_y q_z} \quad \neq \quad \log \frac{p_{xy}}{q_x q_y} + \log \frac{p_{xz}}{q_x q_z} + \log \frac{p_{yz}}{q_y q_z}$$
$$= s_p(x,y) + s_p(x,z) + s_p(y,z)$$

$\Rightarrow$ each sequence is scored as if it descended from the $N-1$ other sequences, instead of a single ancestor.

$\Rightarrow$ evolutionary events are over-counted, problem increases with number of sequences

- **SP scoring error example**

$$\begin{pmatrix} \mathtt{L} & \mathtt{G} & \mathtt{N} & \mathtt{A} \\ \mathtt{L} & \mathtt{N} & \mathtt{A} & \mathtt{G} \\ \mathtt{L} & \mathtt{G} & \mathtt{G} & \mathtt{N} \end{pmatrix} \text{ or } \begin{pmatrix} \mathtt{L} & \mathtt{G} & \mathtt{N} & \mathtt{A} & \mathtt{-} \\ \mathtt{-} & \mathtt{L} & \mathtt{N} & \mathtt{A} & \mathtt{G} \\ \mathtt{-} & \mathtt{L} & \mathtt{G} & \mathtt{G} & \mathtt{N} \end{pmatrix}$$

- How much worse is a G in a conserved L-column compared to completely conserved L-column? Let's see ...

How worse is a G in a conserved L-column?

- Here using BLOSUM90 ($s_p(\mathtt{L},\mathtt{L}) = +5, s_p(\mathtt{G},\mathtt{L}) = -5$)

$$\left.\begin{pmatrix}\mathtt{L}\\ \mathtt{L}\\ \vdots\\ \mathtt{L}\end{pmatrix}\right\} N \text{ versus } N\text{-}1 \left\{\begin{pmatrix}\mathtt{G}\\ \mathtt{L}\\ \vdots\\ \mathtt{L}\end{pmatrix}\right.$$

$$S_c(\underbrace{\mathtt{L},\mathtt{L},\ldots,\mathtt{L}}_{N}) = \sum_{1\le k<l\le N} 5 = \sum_{\ell=1}^{N-1} 5\ell = \frac{5N(N-1)}{2}$$

$$S_c(\mathtt{G},\underbrace{\mathtt{L},\ldots,\mathtt{L}}_{N-1}) = S_c(\mathtt{L},\mathtt{L},\ldots,\mathtt{L}) - 10\,(N-1) = \frac{5N(N\text{-}1)}{2} - 10(N\text{-}1)$$

$s_p(\mathtt{L},\mathtt{L}) - s_p(\mathtt{G},\mathtt{L})$

dependence on inverse N, what does this mean?

- **Fraction:** $\dfrac{S_c(\mathtt{L},\mathtt{L},\ldots,\mathtt{L}) - S_c(\mathtt{G},\mathtt{L},\ldots,\mathtt{L})}{S_c(\mathtt{L},\mathtt{L},\ldots,\mathtt{L})} = \dfrac{10(N-1)}{\frac{5N(N-1)}{2}} = \dfrac{20}{5N} = \dfrac{4}{N}$
- relative difference goes to 0 as $N \to \infty$

SP-error: Consequences

Why is this bad?

1. wrong alignment may be chosen
2. counter-intuitive relative difference should decrease with the amount of evidence for a conserved residue

Why is SP so popular?

- it is all we have
- allows for progressive methods $\Rightarrow$ fast

Goal: Find MSA with best SP over all possible alignments

$\Rightarrow$ keep scoring issues in mind!

Next Question: How to find SP-optimal MSA?

1. exact solution using dynamic programming
   extension of Needleman-Wunsch (2D)
2. improvement of DP algorithm (Carrillo&Lipman, 1988)
3. progressive alignment, e.g. (Feng&Doolittle, 1987)

DP for MSA

- **Method 1:** exact computation by dynamic programming
  $\Rightarrow$ *can be used on any* S_c
- for 2 sequences Needleman&Wunsch: $D_{i,j}$
  = score for best alignment of $a_1^1 \dots a_i^1$ with $a_1^2 \dots a_j^2$.
- for 3 sequences: $D_{i,j,k}$
  = score for best alignment of $a_1^1 \dots a_i^1 (= a)$, $a_1^2 \dots a_j^2 (= b)$ and $a_1^3 \dots a_k^3 (= c)$
- Recursion:

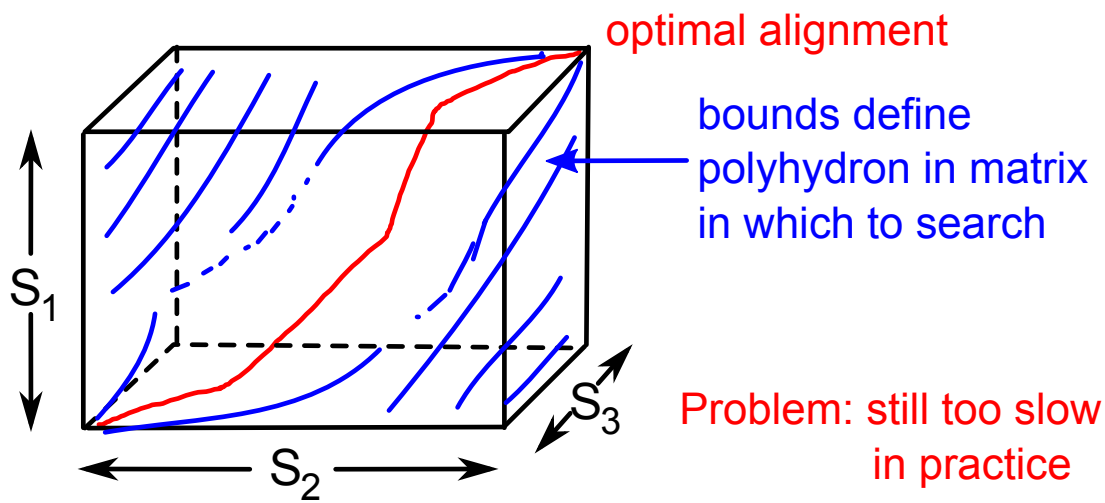
$$D_{i,j,k} = \max \begin{cases} D_{i-1,j-1,k-1} + S_c(a_i, b_j, c_k) & \leftarrow \text{no gap} \\ D_{i-1,j-1,k} + S_c(a_i, b_j, \_) \\ D_{i-1,j,k-1} + S_c(a_i, \_, c_k) \\ D_{i,j-1,k-1} + S_c(\_, b_j, c_k) & \leftarrow \text{one gap} \\ D_{i-1,j,k} + S_c(a_i, \_, \_) \\ D_{i,j-1,k} + S_c(\_, b_j, \_) \\ D_{i,j,k-1} + S_c(\_, \_, c_k) & \leftarrow \text{two gap} \end{cases}$$

- **Remark** three gaps not allowed, i.e. $2^N - 1$ case combinations

DP for MSA of N sequences

- Exponential complexity: $O(n^N)$ time and space
  $\Rightarrow$ need algorithms to reduce search space for SP scoring

  - **Option 1**: reduce search space of DP
    $\Rightarrow$ bounded search via 2D-projection (Carrillo&Lipman, 1988)



  - **Option 2**: progressive alignment

Feng&Doolittle (1987) - Progressive Alignment Idea

Progressive alignment Approach

1. construct guide tree
   - compute all pairwise max. similarity scores
   - calculate distances from similarities $\Rightarrow$ described before
   - generate guide tree $\Rightarrow$ UPGMA $\Rightarrow$ see later
2. generate progressive alignment along guide tree by combining sub-alignments
   - substitute gaps in pairwise alignment with $\diamond$ symbol $\Rightarrow$ aligns with anything for free: $s(\cdot, \diamond) = 0$
   - recompute pairwise alignments with altered sequences $\Rightarrow$ "once a gap, always a gap"
   - best pairwise alignment defines 'fusion' of subalignments

Common application scenario

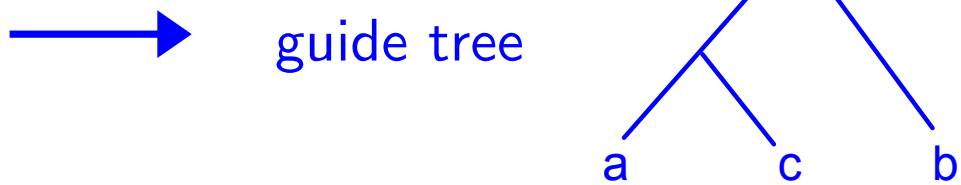
- proteins sequences
- similarities using PAM or BLOSUM
- affine gap penalties

Example (simplified scoring) - Step 1

- 3 sequences: $a = \texttt{ACCAT}$, $b = \texttt{ACGGAT}$, $c = \texttt{AACCAT}$ and score: $s(x,y) = \begin{cases} 0 & \text{if } x \text{ or } y = \diamond \\ 1 & \text{if } x = y \\ -1 & \text{else} \end{cases}$

- pairwise alignments (similarities!):

$a \leftrightarrow b = 2$	AC-CAT ACGGAT
$b \leftrightarrow c = 0$	ACGGAT AACCAT
$a \leftrightarrow c = 4$	-ACCAT AACCAT



Example (continued) - Step 2

- start with $a \leftrightarrow c = 4$ and replace gap by $\diamond$

$$\text{group}_1\text{:} \begin{bmatrix} \diamond\text{ACCAT} \\ \text{AACCAT} \end{bmatrix} \begin{matrix} a' \\ c' \end{matrix}$$

- join $b \Rightarrow$ generate all pairwise alignments from b against group_1

$$S(b, a') = 1 : \begin{bmatrix} \text{-ACGGAT} \\ \diamond\text{AC-CAT} \end{bmatrix} \quad S(b, c') = 1 : \begin{bmatrix} \text{A-CGGAT} \\ \text{AAC-CAT} \end{bmatrix}$$

- use best alignment (b, a') (arbitrary choice) to generate new group

  $\Rightarrow$add gaps in a' from pairwise alignment to all from group_1

$$\text{group}_2\text{:} \begin{bmatrix} \text{-ACGGAT} \\ \diamond\text{AC-CAT} \\ \text{AAC-CAT} \end{bmatrix} \rightarrow \begin{bmatrix} \diamond\text{ACGGAT} \\ \diamond\text{AC}\diamond\text{CAT} \\ \text{AAC}\diamond\text{CAT} \end{bmatrix} \begin{matrix} b'' \\ a'' \\ c'' \end{matrix}$$

- Open question: How to align groups?
  1. consider all alignments for each sequence of group_i with a sequence of group_j
  2. again, best determines combined alignment, i.e. where to put gaps

Remaining Problem

- Conservation of columns is not used

$$\Rightarrow \begin{bmatrix} \texttt{IACL} \\ \texttt{VACI} \end{bmatrix} \text{ and } \begin{bmatrix} \texttt{IACL} \\ \texttt{VACI} \\ \texttt{XACY} \\ \texttt{GACV} \end{bmatrix} \text{ are equally treated}$$

- Improvement idea: progressive alignment with *profiles*
- **Profile:** formally 0-th order Markov chain (= state prob. $\hat{=}$ frequency)

$$\text{alignment} \begin{bmatrix} \texttt{A} & \texttt{A} & \\ \texttt{C} & \texttt{A} & \dots \\ \texttt{C} & \texttt{A} & \\ \texttt{G} & \texttt{A} & \end{bmatrix} \Longrightarrow \quad \text{profile}$$

	c_1	c_2	
A	$\frac{1}{4}$	1	
C	$\frac{1}{2}$	0	...
G	$\frac{1}{4}$	0	
...	0	0	

- **Now:** progressive combination based on pairwise profile alignments, $\Rightarrow$ each profile represents a group (one instead of all-vs-all alignments)
- Several approaches how to score this, but not discussed here ...

Summary

- Multiple Sequence Alignments (MSA) important to identify evolutionary conserved sequence features
- Sum-of-pairs (SP) allows MSA scoring based on pairwise information
- SP scoring overestimates evolutionary events (problematic if many sequences)
- Exact MSA optimization infeasible for more than 5 sequences
- Progressive alignment (PA) allows fast MSA construction via pairwise alignments
- Results can be improved using profile-based scoring in PA