Homology forests revisited

Set of positions in sequence $i$

$$S_{\sigma^i} = \{(i, j) : j \in \{1, \ldots, n_i\}\}$$

Set of positions in all sequences

$$S_{\sigma^1, \ldots, \sigma^k} = \bigcup_{i=1}^{k} S_{\sigma^i}$$

Homology forest

Laminar family over $S_{\sigma^1, \ldots, \sigma^k}$

Rooted forest with leaves labeled by $S_{\sigma^1, \ldots, \sigma^k}$
Alignments today

• Just matching of homologous positions
• No trees, or trees are afterthought
• Homology sets = columns in alignment

CA-GC--TACGGCTTA-GCCT
TA-CCACCTAC--CTGA-GCAT
CA-GCAGTTC--CTTA-GCCT
CA-GC--TACCGGCTGA-ACAT
CATGCAGTTC--CTTACACCT

homologous
Equivalence relations

- **Relation** on a set $S$: a subset of $S \times S$
- We write $a \mathrel{R} b$ if $(a, b)$ is in the relation $R$
- $R$ is an **equivalence relation** if it is:
  - reflexive: $a \mathrel{R} a$ for all $a$ in $S$
  - symmetric: $a \mathrel{R} b \iff b \mathrel{R} a$ for all $a, b$ in $S$
  - transitive: $a \mathrel{R} b, b \mathrel{R} c \implies a \mathrel{R} c$ for all $a, b, c$ in $S$
- Equivalence relations partition $S$ into **equivalence classes**
  - $a \mathrel{R} b \iff a$ and $b$ in the same class, for all $a, b$ in $S$
Alignments = equivalence relations

- We are interested in the relation of Homology ($H$)
- $H$ is an equivalence relation
- Columns of alignment are equivalence classes

CA–GC-- TACGGCTTA–GCCT
TA–CCACTAC--CTGA–GCAT
CA–GCAGTTC--CTTA–GCCT
CA–GC-- TACCGCTGA–ACAT
CATGCAGTTC--CTTACACCT
Alignments we’ve seen

- Motif finding
- Local multiple alignment
- May or may not be actually homologous
- No trees
Order in alignments

• Homology forests do not deal with order of nucleotides in either extant or ancestral species

• However, order is very important for
  • Determining homologous positions
  • Representing and visualizing homology
Global alignment

- Given $k$ homologous sequences
- Assumes sequences are all **colinear**:
  - Homologous positions occur in the same order in each sequence
Partially ordered set

- A partially ordered set (poset) is a set $S$ together with a relation $\leq$ with the following properties (for all $a, b, c \in S$):
  - reflexivity: $a \leq a$
  - antisymmetry: If $a \leq b$ and $b \leq a$, then $a = b$
  - transitivity: If $a \leq b$ and $b \leq c$, then $a \leq c$
Morning tasks poset

• $S = \{\text{wake up, brush hair, shower, leave for school, drink coffee, eat breakfast, pack bag, brush teeth}\}$

• $x \leq y$ if task $y$ may not be done before task $x$
  • shower $\leq$ brush hair
  • eat breakfast $\leq$ brush teeth
  • shower $\leq$ leave for school
  • etc.
• Line upward from a to b if $a \leq b$ and there is no $c \in S$ such that $a \leq c \leq b$

Hasse diagram:

- Leave for school
  - Pack bag
  - Brush teeth
  - Brush hair
  - Drink coffee
  - Eat breakfast
  - Shower
  - Wake up
Partial global multiple alignment

A partial global multiple alignment of sequences $\sigma^1, \ldots, \sigma^k$ is a partially ordered set $P = \{c_1, \ldots, c_m\}$ together with a surjective function $\varphi : S_{\sigma^1, \ldots, \sigma^k} \rightarrow P$ such that $\varphi((i,j_1)) \leq \varphi((i,j_2))$ if $j_1 \leq j_2$

- $c_1, \ldots, c_m$: columns in multiple alignment
- $P$: the “alignment poset”
- surjective: $\varphi$ maps at least one sequence position to every column $c_i$
- $(i,j)$: position $j$ in sequence $i$
Example alignment poset

\[ \varphi \]

(1,1) \rightarrow \sigma_1
(1,2) \rightarrow \sigma_2
(1,3) \rightarrow \sigma_3

\sigma_1: ACT
(2,1) \rightarrow \sigma_1
(2,2) \rightarrow \sigma_2
(2,3) \rightarrow \sigma_3

\sigma_2: AGT
(2,1) \rightarrow \sigma_1
(2,2) \rightarrow \sigma_2
(2,3) \rightarrow \sigma_3

\sigma_3: TGT
(2,1) \rightarrow \sigma_1
(2,2) \rightarrow \sigma_2
(2,3) \rightarrow \sigma_3

(3,1) \rightarrow \sigma_5
(3,2) \rightarrow \sigma_4
(3,3) \rightarrow \sigma_3

• Unless we have a **total order**, the order of some columns is not specified (e.g., \(c_5\) and \(c_1\))

one possible alignment:

\[ \begin{align*}
\sigma_1: & \quad \text{ACT} \\ (2,1) & \rightarrow c_1 \\
(2,2) & \rightarrow c_4 \\
(2,3) & \rightarrow c_3 \\
(3,1) & \rightarrow c_5 \\
(3,2) & \rightarrow c_4 \\
(3,3) & \rightarrow c_3 \\
\end{align*} \]

\[ \begin{align*}
\sigma_2: & \quad \text{AGT} \\ (2,2) & \rightarrow c_4 \\
(2,3) & \rightarrow c_3 \\
(3,2) & \rightarrow c_4 \\
(3,3) & \rightarrow c_3 \\
\end{align*} \]

\[ \begin{align*}
\sigma_3: & \quad \text{TGT} \\ (2,3) & \rightarrow c_3 \\
(3,1) & \rightarrow c_5 \\
(3,2) & \rightarrow c_4 \\
(3,3) & \rightarrow c_3 \\
\end{align*} \]

\[ c_1 \leq c_2 \leq c_3 \\
\sigma_5 \leq c_4 \leq c_3 \\
\sigma_1 \leq c_4 \leq c_3 \\
\]
Extreme example: Null alignment

- Given sequences $\sigma^1, \sigma^2, \ldots, \sigma^k$ of lengths $n_1, n_2, \ldots, n_k$

- Null alignment: size of $P$ is $\sum n_i$

- Every position is mapped to a different column

```
ACT------  ------ACT  A---CT---
---AGT---  or AGT------  or -A-G--T--  or...
-------TGT  ----TGT----  --T-----GT
```
Number of pairwise partial global alignments

Proposition 2.5  The number of partial global alignments of two sequences of length $n$ and $m$ is \( \binom{n+m}{m} \).

Proof: Note that the number of alignments with $k$ homologous nucleotides is given by \( \binom{n}{k} \binom{m}{k} \). The total number of alignments is therefore

\[
\sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} = \binom{n+m}{n}.
\]
Linear extension

- **A linear extension** of a partially ordered set $P = \{c_1, \ldots, c_m\}$ is a permutation $\Pi$ of the elements $c_1, \ldots, c_m$ such that whenever $c_i \leq c_j$, $\Pi(c_i) \leq \Pi(c_j)$.

- **A global multiple alignment** is a partial global multiple alignment along with a linear extension of the alignment poset.

\[
\begin{align*}
c_1 \leq & \quad c_2 \leq \quad c_3 \\
c_1 \leq & \quad c_4 \leq \quad c_3 \\
c_5 \leq & \quad c_4 \leq \quad c_3
\end{align*}
\]

linear extension

\[
\begin{align*}
\text{linear extension} & \quad C_1 \quad C_5 \quad C_2 \quad C_4 \quad C_3 \\
A & - \quad C \quad - \quad T \\
A & - \quad - \quad G \quad T \\
- \quad T \quad - \quad G \quad T
\end{align*}
\]
The number of pairwise global alignments of sequences of length $n$ and $m$ is the Delannoy number $D_{n,m}$.

$D_{n,m}$: number of lattice paths from $(0,0)$ to $(n,m)$ with allowed moves $\uparrow, \rightarrow$, and $\searrow$.

$$D_{n,m} = \min(n,m)\sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} 2^k$$
Partial vs full alignments

- The number of possible partial alignments is very large
- The number of full alignments is even larger

<table>
<thead>
<tr>
<th>n</th>
<th>partial</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>321</td>
</tr>
<tr>
<td>5</td>
<td>252</td>
<td>1,683</td>
</tr>
<tr>
<td>6</td>
<td>924</td>
<td>8,989</td>
</tr>
<tr>
<td>7</td>
<td>3,432</td>
<td>48,639</td>
</tr>
<tr>
<td>8</td>
<td>12,870</td>
<td>265,729</td>
</tr>
<tr>
<td>9</td>
<td>48,620</td>
<td>1,462,563</td>
</tr>
<tr>
<td>10</td>
<td>184,756</td>
<td>8,097,453</td>
</tr>
</tbody>
</table>
Representing full pairwise global alignments

- representations for pairwise alignment $h$, of sequences $\sigma^1 = \sigma^1_1 \sigma^1_2 \cdots \sigma^1_n$ and $\sigma^2 = \sigma^2_1 \sigma^2_2 \cdots \sigma^2_m$
- edit string $h$ over edit alphabet $\{H, I, D\}$
  - $H$: homology, $I$: insertion, $D$: deletion
- path in alignment graph
- sequence of pairs $(\sigma^1_i \diamond \sigma^2_j), (\sigma^1_i \diamond -), (\sigma^2_j \diamond -)$
Comparing alignments

• Compare in terms of H, I, and D pairs:

\[ h_H = \{(i, j) : (\sigma_i^1 \diamond \sigma_j^2) \in h\}, \]

\[ h_D = \{i : (\sigma_i^1 \diamond -) \in h\}, \]

\[ h_I = \{j : (\sigma_j^2 \diamond -) \in h\}. \]

for any \( h \in \mathcal{A}_{n,m} \quad |h_H| + |h_D| = n \text{ and } |h_H| + |h_I| = m. \)
Alignment equivalence

• Alignments are defined to be equivalent if they match up the same sequence positions

\[ h^i \sim h^j \text{ if and only if } h^i_H = h^j_H \]

• Equivalently, alignments are equivalent if they gap the same sequence positions

\[ h^i \sim h^j \text{ if and only if } h^i I = h^j I \text{ and } h^i D = h^j D \]

\[
\begin{align*}
\text{AC-T} & \sim \text{A-CT} \\
\text{A-GT} & \sim \text{AG-T}
\end{align*}
\]
Alignment measures

• Measures of sensitivity and specificity, with respect to a reference alignment $h^r$

\[
\begin{align*}
  f(h^i, h^j) &= \frac{|h^i_H \cap h^j_H|}{|h^i_H|} \\
  f_D(h^p, h^r) &= f(h^r, h^p) = \frac{|h^r_H \cap h^p_H|}{|h^r_H|} \\
  f_M(h^p, h^r) &= f(h^p, h^r) = \frac{|h^r_H \cap h^p_H|}{|h^p_H|}
\end{align*}
\]

“developer’s measure”
(sensitivity)

“modeler’s measure”
(specificity)
Distances between alignments

• Would like a distance function between alignments

• Should be a metric, i.e., it should satisfy:

\[ d(h^i, h^j) \geq 0 \quad \forall h^i, h^j \in A_{n,m}, \]

\[ d(h^i, h^j) = 0 \text{ iff } h^i \sim h^j \quad \forall h^i, h^j \in A_{n,m}, \]

\[ d(h^i, h^j) = d(h^j, h^i) \quad \forall h^i, h^j \in A_{n,m}, \]

\[ d(h^i, h^j) + d(h^j, h^k) \geq d(h^i, h^k) \quad \forall h^i, h^j, h^k \in A_{n,m}. \]
Alignment metric

- The following function is a finite metric on alignments:

\[
  d(h^i, h^j) = 2|h^i_H| + |h^i_I| + |h^i_D| - 2|h^i_H \cap h^j_H| \\
  - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D| \\
  = 2|h^j_H| + |h^j_I| + |h^j_D| - 2|h^i_H \cap h^j_H| \\
  - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D| \\
  = n + m - 2|h^i_H \cap h^j_H| \\
  - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D|).
\]

(Schwartz & Pachter, 2007)
Example alignment metric

Metric for $A_{2,2}$

<table>
<thead>
<tr>
<th></th>
<th>$HH$</th>
<th>$HDI$</th>
<th>$DIH$</th>
<th>$IHD$</th>
<th>$DHI$</th>
<th>$DDII$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$HH$</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$HDI$</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$DIH$</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$IHD$</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$DHI$</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$DDII$</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

$$d(h^i, h^j) = n + m - 2|h^i_H \cap h^j_H|$$

$$- |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D|).$$
Alignment metric accuracy

- Instead of developer or modeler score, use score based on metric

\[ g(h^i, h^j) = 1 - \frac{d(h^i, h^j)}{n+m} \]

- \( g \) is fraction of positions aligned identically in the two alignments

- Alignment Metric Accuracy (AMA) = \( g(h^p, h^r) \)

(Schwartz & Pachter, 2007)
Multiple alignment accuracy

• All multiple alignments of sequences of lengths \(n_1, n_2, \ldots, n_k\): \(A_{n_1,n_2,...,n_k}\)

Given two MSAs \(h^i, h^j \in A_{n_1,n_2,...,n_k}\):

\[
d(h^i, h^j) = \sum_{s^1=1}^{k-1} \sum_{s^2>s^1}^{k} d(h^i_{s^1,s^2}, h^j_{s^1,s^2})
\]

• Like sum-of-pairs scoring

• Accuracy:

\[
g(h^p, h^r) = 1 - \frac{d(h^p, h^r)}{(k-1) \sum_{i=1}^{k} n_i}.
\]

(Schwartz & Pachter, 2007)