References

Multiples Sequenzalignment (MSA)

Sequence Family

Subtyp A1: GAGCAGAAAGACAGGGGAACAGGCCCAACCCTTAGTT
Subtyp A2: GAGAACAGGGAGCCGTCACCACCCCTGCAATT
Subtyp B: GAGCCGATAGACAAGGAATGTATCCTTTAACT
Subtyp C: GAGACGATAGACAAGGAACTGCCCTTTAACT
Subtyp D: GAGCAAGAAGACAAGGAACTGTATCCTTTAACT
Subtyp F1: GAGCAGAAGGGAGGACAGGGACTGTATCCTCCCTTAGCC
Subtyp G: GAGCAGAAGGGAAAGGAACTATATCCTCTATCT
Subtyp H: GAGCAGCTGAAGGACAAGGAACCTCCCTTAGCT
Subtyp J: GAGCCGAAGGACAAGGAACTGTATCCTCTAACT
Subtyp K: GAGACCAAAGACAAGGAACAGGCCCTCCTTTAACT

some segments of HIV sequences from a genic region

Hypotheses:
- all sequences are derived from a single ancestral sequence
- the sequence changes during evolution by mutations, insertions and deletions
# Multiples Sequenzalignment (MSA)

## Sequence Family

<table>
<thead>
<tr>
<th>Subtyp</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>GAGCA---GAAAGACAG------GGAACAGGCCCAACCCTTAGTT</td>
</tr>
<tr>
<td>A2</td>
<td>GAGAA------CAG------GGAGCCGTCCACCCTGCAATT</td>
</tr>
<tr>
<td>B</td>
<td>GAGCC---GATAGACAA------GGAACTGTA---TCCTTTAACT</td>
</tr>
<tr>
<td>C</td>
<td>GAGAC---GATAGACAA------GGAACT------GCCCTTAACT</td>
</tr>
<tr>
<td>D</td>
<td>GAGCA---GAAAGACAA------GGAACTGTA---TCCTTTAACT</td>
</tr>
<tr>
<td>F1</td>
<td>GAGCA---GAAGGACGAGGGACAGGGACTGTATCCTCTCTAGCC</td>
</tr>
<tr>
<td>G</td>
<td>GAGCA---GAAGGAAAA------GGAACTATA---TCCTCTATCT</td>
</tr>
<tr>
<td>H</td>
<td>GAGCAGCTGAAGGACAA------GGAACC------TCCCTAGCT</td>
</tr>
<tr>
<td>J</td>
<td>GAGCC---GAAGGACAA------GGAACTGTA---TCCTCTAACT</td>
</tr>
<tr>
<td>K</td>
<td>GAGAC---CAAAGACAA------GGAACAGAGCCCTTCTTTAACT</td>
</tr>
</tbody>
</table>

**Alignment of the sequences:**

Gap characters ‘-’ are introduced, such that positions that stem from the same ancestral position are underneath each other.
PFAM

• database of currently ~13,000 protein sequence families
• represented as MSAs and so-called profile HMMs
• some families have >100,000 members
• protein sequence families of PFAM found in ~80% of proteins

Example (protein domain Hexokinase_2 (PF03727))

<table>
<thead>
<tr>
<th>Pfam seed alignment for PF03727</th>
</tr>
</thead>
<tbody>
<tr>
<td>P33284/224-472 QTKMGIIGTGSGAYDVWSGIEKLEG...EDIGDSPPAINCYEYGF...DNEHLV...P1KYDVIIDEE...SPPPGOQAFERITSGYLLGEIMRLVLLD</td>
</tr>
<tr>
<td>P50521/213-449 DTFKGIIFGTGSGAYDESNGPQKLAK...CTGDHALMIVGAT...DFSSLHNL...YDILLHHD...TPNAQRIFEKRVGGNLGEIRFLRH</td>
</tr>
<tr>
<td>P3555/219-458 QCEGVGIVGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P52722/223-462 NCEGVGIVGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P27725/223-462 HCEGVGIVGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P27725/223-462 HCEGVGIVGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P17706/230-500 FVYGCGFSGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P50521/224-469 EAKMGKFSGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P04807/225-473 ETKMGKFSGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P27725/671-910 TCEGVGIVGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>Q09756/223-469 GTEIGVFSGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P04806/225-472 ETKMGKFSGTGSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P52790/236-475 PCEVGLVGGGGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P52790/671-910 RCEVGLVGGGGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>Q05215/239-465 PCYIGILGTSGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
<tr>
<td>P52789/671-910 HCEVGLVGGGGAYDEEMNKLVE...DGRCLCVWGAEGDSEGDFLLELYDLVDES...SANPQOKLYKLVolume: LRLVLLRL</td>
</tr>
</tbody>
</table>
Deciding for Membership in a Sequence Family

### Membership

- **Want to know:** Given sequence family $F$ and protein sequence $S$, is $S$ a member of $F$?
- **Want to search large-scale:**
  - a single family $F$ against a large set of protein sequences $S_1, S_2, \ldots$ (e.g. all proteins of 1 species)
  - a single protein sequence $S$ against many families $F_1, F_2, \ldots$ (e.g. all PFAM families)

### Remote Homology

Want to

- uncover remote relatedness: low conservation, many insertions, deletions
- be more sensitive than BLAST database searching
Standard Alignment not Well Suited

Example

Pfam seed alignment for PF03727

Disadvantages of BLAST and MSA scoring methods

- conservation and frequencies of insertions, deletions not position-specific
- theoretical limit of pairwise scoring (scoring-matrix): does not capture all prior information about conservation
Profile Hidden Markov Models

### Profile HMM

- introduced by Krogh 1994 and Eddy 1995
- typically done for sequences of amino acids
- PFAM protein sequence families are stored as profile HMMs
- popular tool: HMMer ("hammer")
  - highly optimized version HMMer 3
  - searches 10 million protein sequences against one profile in 5 min
Profile HMM

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>C A - - - G A A A</td>
</tr>
<tr>
<td>A2</td>
<td>A A - - - - - - -</td>
</tr>
<tr>
<td>B</td>
<td>C C - - - G A T A</td>
</tr>
<tr>
<td>C</td>
<td>A C - - - G A T A</td>
</tr>
<tr>
<td>D</td>
<td>C A - - - G A A A</td>
</tr>
<tr>
<td>F1</td>
<td>C A - - - G A A G</td>
</tr>
<tr>
<td>G</td>
<td>C A - - - G A A G</td>
</tr>
<tr>
<td>H</td>
<td>C A G C T G A A G</td>
</tr>
<tr>
<td>J</td>
<td>C C - - - C A A A</td>
</tr>
<tr>
<td>K</td>
<td>A C - - - C A A A</td>
</tr>
</tbody>
</table>

Match states $M_i$ for every sufficiently conserved position, emits 1 character
Insert states $I_i$ to allow for insertions, emits 1 character
Delete states $D_i$ to allow for deletions, emits $\varepsilon$
A profile HMM is a GHMM

Example (from MSA on previous slide)

set of states $Q = \{B, M_1, M_2, \ldots, M_6, I_0, \ldots, I_6, D_1, \ldots, D_6, E\}$

emission $y = y_1y_2 \cdots y_n$ is a (DNA) sequence
parse $x = ((q_1, v_1), (q_2, v_2), \ldots, (q_t, v_t))$ ($q_i \in Q$, $0 =: v_0 \leq v_1 \leq \cdots \leq v_t = n$)

Here, all Delete states are silent, all Match and Insert states emit 1 character: $P(x, y) > 0$ implies

\[
\begin{align*}
    v_i - v_{i-1} = 0 & \iff q_i \in \{D_1, D_2, \ldots\} \\
    v_i - v_{i-1} = 1 & \iff q_i \in \{M_1, M_2, \ldots, I_0, I_1, \ldots\}
\end{align*}
\]

Therefore, $x$ is already specified by state sequence $(q_1, q_2, \ldots, q_t)$. 
A profile HMM is a GHMM

Example (from MSA on previous slide)

set of states $Q = \{ B, M_1, M_2, \ldots, M_6, l_0, \ldots, l_6, D_1, \ldots, D_6, E \}$

emission $y = y_1 y_2 \cdots y_n = a$ (DNA) sequence

parse $x = ((q_1, v_1), (q_2, v_2), \ldots, (q_t, v_t))$ ( $q_i \in Q, 0 =: v_0 \leq v_1 \leq \cdots \leq v_t = n$).

Example:

$x = (B, M_1, M_2, l_2, l_2, M_3, D_4, D_5, M_6, E)$

$y = \text{CATCGG}$

$$P(x, y) = P_{\text{trans}}(M_1 | B) \cdot P_{\text{emi}}(C | M_1) \cdot P_{\text{trans}}(M_2 | M_1) \cdot P_{\text{emi}}(A | M_2)$$

$$\cdot P_{\text{trans}}(l_2 | M_2) \cdot P_{\text{emi}}(T | l_2) \cdot P_{\text{trans}}(l_2 | l_2) \cdot P_{\text{emi}}(C | l_2)$$

$$\cdot P_{\text{trans}}(M_3 | l_2) \cdot P_{\text{emi}}(G | M_3) \cdot P_{\text{trans}}(D_4 | M_3) \cdot P_{\text{trans}}(D_5 | D_4)$$

$$\cdot P_{\text{trans}}(M_6 | D_5) \cdot P_{\text{emi}}(G | M_6) \cdot P_{\text{trans}}(E | M_6)$$
Parse (B, M₁, M₂, I₂, I₂, M₃, D₄, D₅, M₆, E) of sequence y = CATCGG defines sequence-to-profile alignment:

\[
\begin{array}{cccccccc}
M₁ & M₂ & - & - & M₃ & M₄ & M₅ & M₆ \\
C & A & T & C & G & - & - & G \\
\end{array}
\]

(approximately in agreement with the MSA)
Is protein $S$ a member of profile $F$?

1. build profile HMM $P_F(x, y)$ for sequence family $F$
2. compute the probability of $S$ in this model
   \[ P_F(S) = \sum_{\text{parse } x} P_F(x, S) \]
   with the Forward algorithm
3. decide: $S$ is member of $F$ iff
   \[ P_F(S) \geq t_F \]
   for some threshold $t_F$ specific to $F$
### Estimation of Parameters

**Need to estimate...**

- **emission probabilities** $P_{emi}(\cdot|q)$ for Match and Insert states $q$
- **transition probabilities** $P_{trans}(q|q')$

in order to reflect

- typical or required characters at match columns and insert regions
- position-specific likelihood of insertions and deletions for the given sequence family.
The MSA induces **observed** frequencies of emissions and transitions.

E.g.:

In $M_1$ we observe emissions: 3xA, 7xC
In $I_2$ we observe emissions: 1xC, 1xG, 1xT
transitions from $M_2$: 8x to $M_3$, 1x to $D_3$, 1x to $I_2$
Estimation of the Parameters of a Multinomial Distribution

Parameters of Multinomial Distribution

General Task:

A random experiment has \( k \) different outcomes \( 1, 2, \ldots, k \) for which we want to estimate probabilities \( p_1, p_2, \ldots, p_k \).

We have

1. a vector of frequencies \( \vec{n} = (n_1, n_2, \ldots, n_k) \)  
(sample of size \( n := n_1 + \cdots + n_k \))

2. prior knowledge on the distribution of \( \vec{p} = (p_1, p_2, \ldots, p_k) \)

Example

Ex 1: Travel to unknown country. First two locals have black hair. \( 1= \)black hair, \( 2= \)not black hair. \( \vec{n} = (2, 0) \). Seek \( p_1 \sim \) fraction of locals with black hair.

Ex 2: Normal Dice. Each side is labeled 1, 2 or 3. Throw 7 times. \( \vec{n} = (4, 3, 0) \). Seek probabilities for 1, 2, 3.
Simple Methods to estimate $\hat{\rho}$

Relative frequencies

$$\hat{p}_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$$

**Disadvantage:** Unobserved outcomes obtain a probability of 0: $\hat{p}_i = 0$ if $n_i = 0$.

Relative frequencies with pseudocounts

$$\hat{p}_i = \frac{n_i + c_i}{\sum_{j=1}^{k} (n_j + c_j)}$$

c$_i$ is called pseudocount for outcome $i$, e.g. $c_i = 1 \ \forall i$. The effect of the pseudocounts decreases with increasing sample size.

**Disadvantage:** Similarity between amino acids is not considered.
Simple Methods to estimate $\bar{p}$

Using a scoring matrix

Suppose $\bar{n}$ is the counts of the 20 amino acids in an alignment column, $p(i,j)$, $p(j)$ are the probability of observing the aligned pair $(i,j)$ or $j$ respectively in a large set of aligned sequences. Let

$$s(i,j) := \log_2 \frac{p(i,j)}{p(i)p(j)}$$

be the corresponding score of $i$ with $j$ (given in a scoring matrix).

We could define

$$\hat{p}_i \sim \sum_j p(i | j) \cdot n_j \sim p(i) \sum_j 2^{s(i,j)} \cdot n_j,$$

and normalize such that $\sum_i \hat{p}_i = 1$. This approach considers some prior knowledge: pairwise similarity of amino acids.

Disadvantages:

- When $n_i \to \infty$ and $n_j = 0$ for $j \neq i$, then $\hat{p}_i \not\to 1$ in general.
- Pairwise similarity does not fully capture prior knowledge.
## Intuition

A phenylalanine (F) is an aromatic, large non-polar residue. In one context an F can therefore play the role of an aromatic residue, in another context it can play the role of a large non-polar residue.

A scoring matrix cannot distinguish between the two roles. Ideally, the score of an amino acid against a column of the MSA should depend on the combination of residues in the column.
Bayesian Model

In profile HMMs usually a mean posterior estimate in a Bayesian model is used to estimate emission distributions:

\[ \hat{\rho}_i := E[p_i | \vec{n}] \]

Prior knowledge is given by a distribution on the set of all probability vectors \( \vec{p} \) by a density

\[ f(\vec{p}) \]

that is estimated on a large set of alignment columns.

Reference