Hidden Markov models for sequence alignment

Niko Beerenwinkel

Outline

- Pair HMMs
- Pairwise sequence alignment
- Profile HMMs
- Multiple sequence alignment

Global alignment

- Problem:

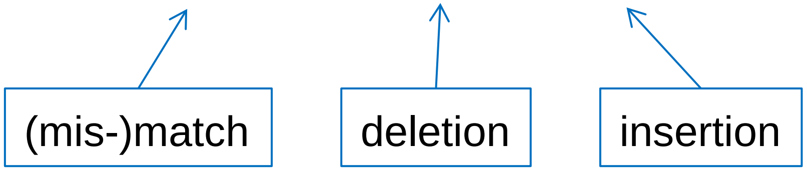
  Given two (DNA or protein) sequences, which characters have descended from a common ancestor?

- For example,

$$x = \texttt{HEAGAWGHEE}$$
$$y = \texttt{PAWHEAE}$$

Global alignment

$$\begin{aligned} x &= \mathtt{HEAGAWGHE{-}E} \\ y &= \mathtt{{-}{-}P{-}AW{-}HEAE} \end{aligned}$$



(mis-)match

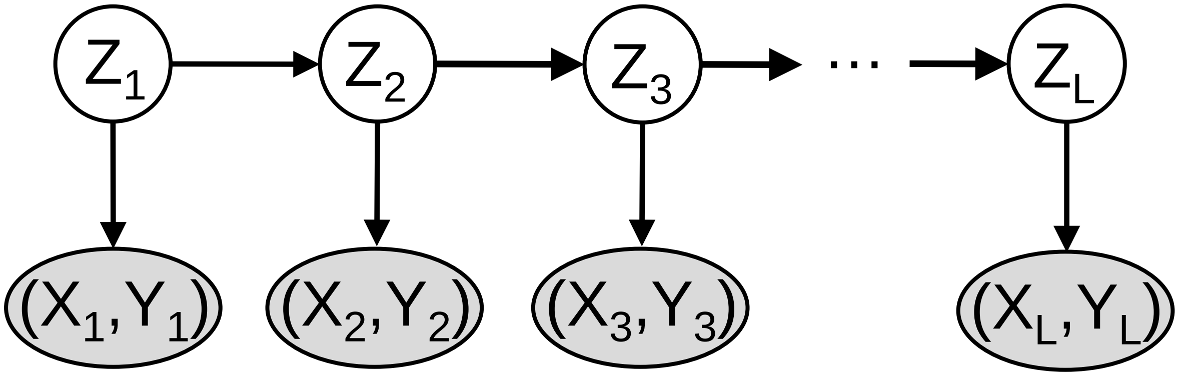
deletion

insertion

- We think of a (global) alignment as a probabilistically generated sequence of *pairs* of symbols.
- All pairs except (−,−) are allowed.

The pair HMM for global alignment

$$
\begin{aligned}
z &= \texttt{XXMXMMXMMYM} \\
x &= \texttt{HEAGAWGHE-E} \\
y &= \texttt{--P-AW-HEAE}
\end{aligned}
$$



Emission probabilities

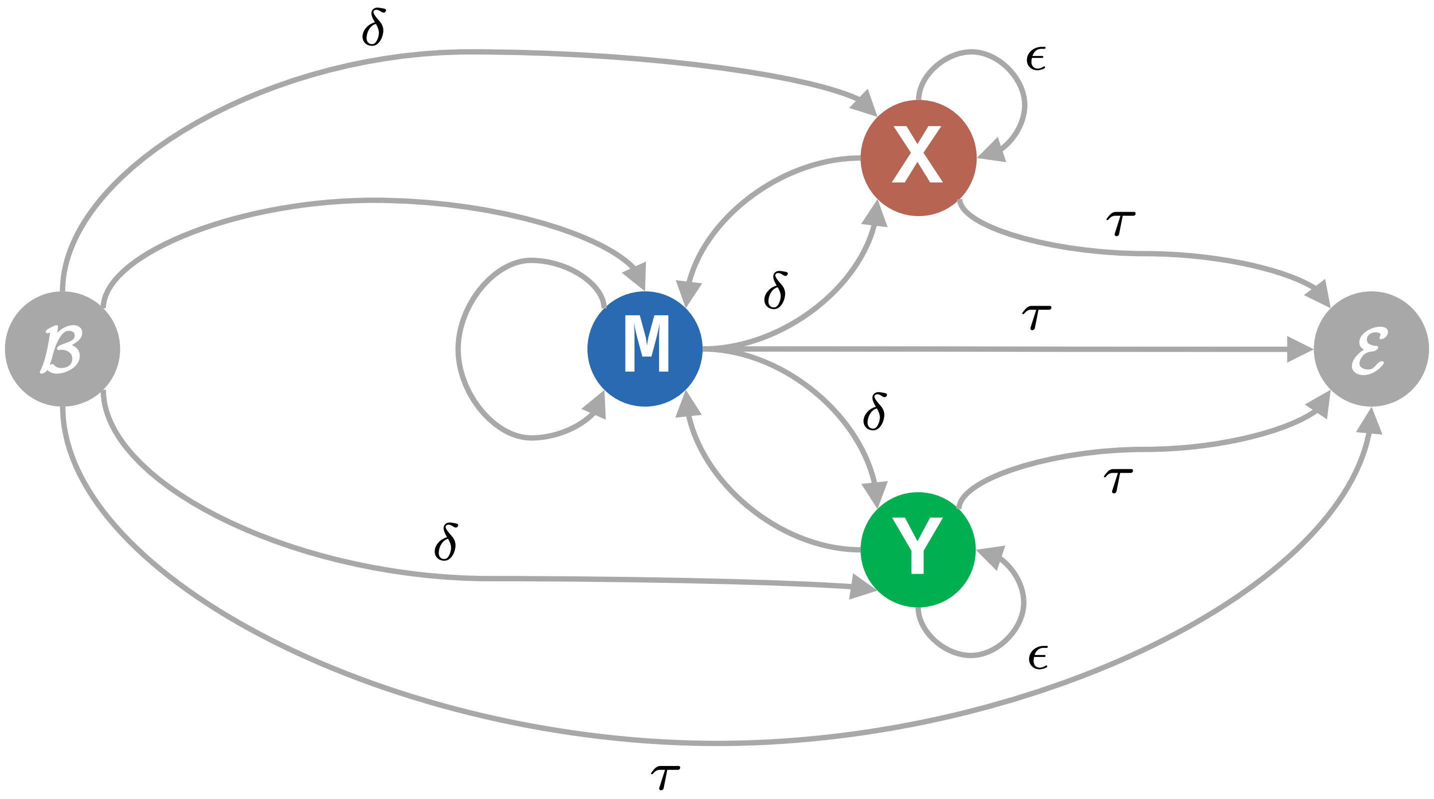
$$\begin{aligned} z &= \texttt{XXMXMMXMMYM} \\ x &= \texttt{HEAGAWGHE-E} \\ y &= \texttt{--P-AW-HEAE} \end{aligned}$$

$$\begin{aligned} P[(X,Y)=(x_i,y_j) \mid Z=\texttt{M}] &= E_{\texttt{M},(x_i,y_j)} = p_{x_i,y_j} \\ P[(X,Y)=(x_i,-) \mid Z=\texttt{X}] &= E_{\texttt{X},(x_i,-)} = q_{x_i} \\ P[(X,Y)=(-,y_j) \mid Z=\texttt{Y}] &= E_{\texttt{Y},(-,y_j)} = q_{y_j} \end{aligned}$$

Notation

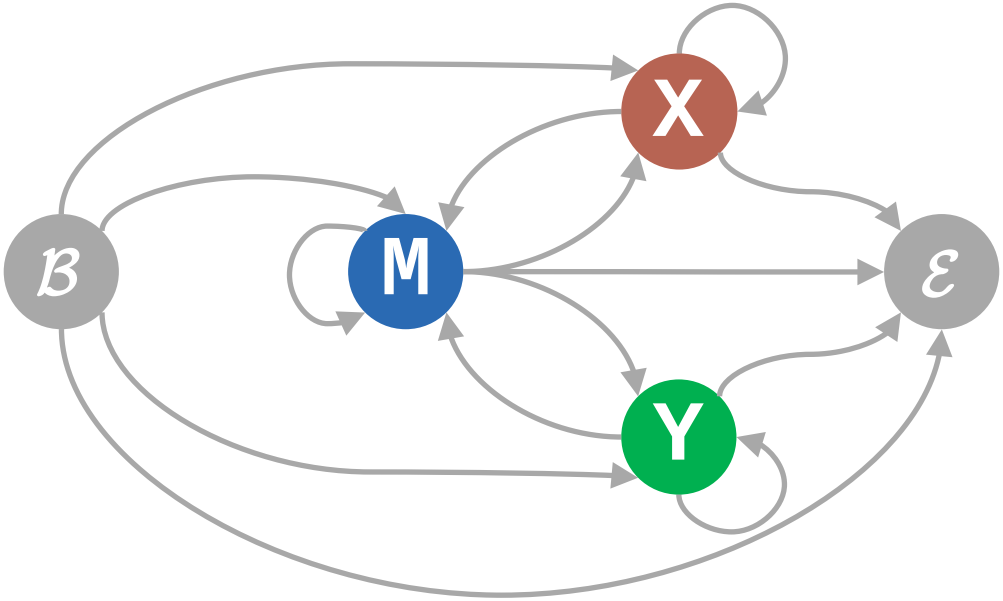
$$\begin{aligned} z &= (z_1, \ldots, z_L) \\ x &= (x_1, \ldots, x_n) = (x_i)_{i=1,\ldots,n} \\ y &= (y_1, \ldots, y_m) = (y_j)_{j=1,\ldots,m} \end{aligned}$$

State space



$\mathcal{B}$ and $\mathcal{E}$ are *silent* states.

Transition probabilities



$$(P(Z_i \mid Z_{i-1})) = \begin{array}{c} \\ \mathcal{B} \\ \mathtt{M} \\ \mathtt{X} \\ \mathtt{Y} \\ \mathcal{E} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathcal{B} & \mathtt{M} & \mathtt{X} & \mathtt{Y} & \mathcal{E} \end{array} \\ \begin{pmatrix} 0 & * & \delta & \delta & \tau \\ 0 & * & \delta & \delta & \tau \\ 0 & * & \epsilon & 0 & \tau \\ 0 & * & 0 & \epsilon & \tau \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Remark: Local alignment is similar by flanking the global model with an additional random model at the beginning and the end.

Optimal alignments

- The most probable state path of the pair HMM is the optimal alignment.
- We have to compute

$$z^* = \underset{z}{\text{argmax}}\, P(x, y, z)$$

→ Viterbi algorithm!

- Let $v^M(i, j)$ be the probability of the most probable path emitting (x_i, y_j) in state M, and similarly for v^X, v^Y, and $v^{\mathcal{E}}$.
- Then $v^{\mathcal{E}} = P(x, y, z^*)$.
- For simplicity, we assume that the begin state is M.

Viterbi algorithm for pair HMMs

- **Initialization**: $v^M(0,0) = 1$, else $v^*(i,0) = v^*(0,j) = 0$
- **Recurrence**: for i = 1, …, n and j = 1, …, m:

$$v^M(i,j) = p_{x_i,y_j} \max \begin{cases} (1-2\delta-\tau)v^M(i-1,j-1) \\ (1-\epsilon-\tau)v^X(i-1,j-1) \\ (1-\epsilon-\tau)v^Y(i-1,j-1) \end{cases}$$

$$v^X(i,j) = q_{x_i} \max \begin{cases} \delta v^M(i-1,j) \\ \epsilon v^X(i-1,j) \end{cases}$$

$$v^Y(i,j) = q_{y_j} \max \begin{cases} \delta v^M(i,j-1) \\ \epsilon v^Y(i,j-1) \end{cases}$$

- **Termination**: $v^{\mathcal{E}} = \tau \max \left\{ v^M(n,m),\ v^X(n,m),\ v^Y(n,m) \right\}$

The probability of two sequences being related

- The joint probability of x and y, *irrespective of their alignment* z, is

$$P(x,y) = \sum_{\text{alignments } z} P(x,y,z)$$

→ Forward algorithm!

- Write $x_i \diamond y_j$ if characters x_i and y_j are aligned.
- Let $f^M(i, j) = P(x_1,\dots,x_i, y_1,\dots,y_j, x_i \diamond y_j)$ be the joint probability of the subsequences and $x_i \diamond y_j$, and similarly for f^X, f^Y, $f^{\mathcal{E}}$.
- Then $f^{\mathcal{E}}(n, m) = P(x, y)$.

Forward algorithm for pair HMMs

- **Initialization**: $f^M(0,0) = 1,\ f^X(0,0) = f^Y(0,0) = 0,$
  and all $f^*(i,-1) = f^*(-1,j) = 0$
- **Recurrence**: for i = 0, …, n and j = 1, …, m, except (0,0):

$$\begin{aligned} f^M(i,j) &= p_{x_i,y_j}\Big\{(1-2\delta-\tau)f^M(i-1,j-1)+ \\ &\qquad +(1-\epsilon-\tau)\left[f^X(i-1,j-1)+f^Y(i-1,j-1)\right]\Big\} \end{aligned}$$

$$f^X(i,j) = q_{x_i}\left\{\delta f^M(i-1,j)+\epsilon f^X(i-1,j)\right\}$$

$$f^Y(i,j) = q_{y_j}\left\{\delta v^M(i,j-1)+\epsilon f^Y(i,j-1)\right\}$$

- **Termination**: $f^{\mathcal{E}}(n,m) = \tau\left\{f^M(n,m)+f^X(n,m)+f^Y(n,m)\right\}$

Full probability versus Viterbi path

- The posterior of an alignment is

$$P(z \mid x, y) = \frac{P(x, y, z)}{P(x, y)}$$

- In particular, the probability of the Viterbi path is

$$P(z^* \mid x, y) = \frac{v^{\mathcal{E}}(n, m)}{f^{\mathcal{E}}(n, m)}$$

- In general, this probability is very small, because many equally (or almost equally) good alignments exist.
- Therefore, P(x, y) is usually more meaningful than P(x, y, z*).

Example: hemoglobin

```
x = HBA_HUMAN  KVADALTNAVAHVD-----DMPNALSALSDLH
               KV   + +A  ++          +L+ L+++H
y = LGB2_LUPLU KVFKLVYEAAIQLQVTGVVVTDATLKNLGSVH

    HBA_HUMAN  KVADALTNAVAHVDDM-----PNALSALSDLH
               KV   + +A  ++          +L+ L+++H
    LGB2_LUPLU KVFKLVYEAAIQLQVTGVVVTDATLKNLGSVH

    HBA_HUMAN  KVADALTNA-----VAHVDDMPNALSALSDLH
               KV   + +A     V  V     +L+ L+++H
    LGB2_LUPLU KVFKLVYEAAIQLQVTGVVVTDATLKNLGSVH
```

$$P(x, y, z^*) = 4.6 \times 10^{-6}$$

Local accuracy: the posterior of $x_i \diamond y_j$

- We want to compute

$$P(x_i \diamond y_j \mid x, y) = \frac{P(x_i \diamond y_j, x, y)}{P(x, y)}$$

- The joint probability in the numerator is

$$P(x, y, x_i \diamond y_j) = P(x_{1\ldots i}, y_{1\ldots j}, x_i \diamond y_j) \cdot$$
$$\cdot \; P(x_{(i+1)\ldots n}, y_{(j+1)\ldots m} \mid x_i \diamond y_j)$$
$$= f^M(i,j)\, b^M(i,j)$$

→ Backward algorithm!

Backward algorithm for pair HMMs

- **Initialization**: $b^M(n,m) = b^X(n,m) = b^Y(n,m) = \tau$,
  and all $b^*(i, m+1) = b^*(n+1, j) = 0$
- **Recurrence**: for i = n, …, 1 and j = m, …, 1, except (n,m):

$$\begin{aligned}
b^M(i,j) &= (1-2\delta-\tau)p_{x_{i+1},y_{j+1}}b^M(i+1,j+1) + \\
&\quad +\delta\left[q_{x_{i+1}}b^X(i+1,j) + q_{y_{j+1}}b^Y(i,j+1)\right]
\end{aligned}$$

$$\begin{aligned}
b^X(i,j) &= (1-\epsilon-\tau)p_{x_{i+1},y_{j+1}}b^M(i+1,j+1) + \\
&\quad +\epsilon q_{x_{i+1}}b^X(i+1,j)
\end{aligned}$$

$$\begin{aligned}
b^Y(i,j) &= (1-\epsilon-\tau)p_{x_{i+1},y_{j+1}}b^M(i+1,j+1) + \\
&\quad +\epsilon q_{y_{j+1}}b^Y(i,j+1)
\end{aligned}$$

Alignment accuracy

- The expected number of correctly aligned characters in an alignment z is

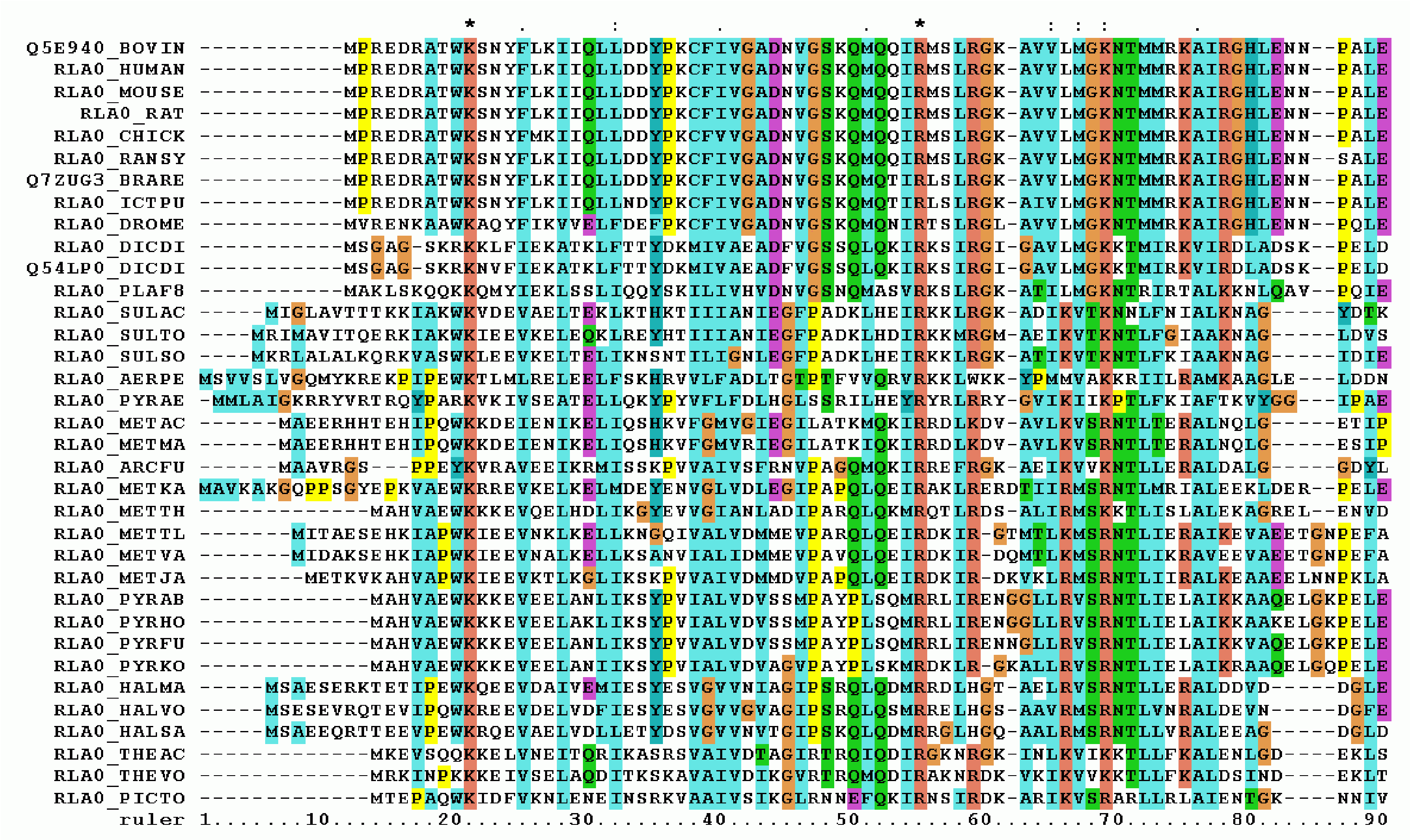
$$A(z) := \sum_{x_i \diamond y_j \text{ in } z} P(x_i \diamond y_j)$$

- Another dynamic program maximizes this score,

$$A(i,j) = \max \begin{cases} A(i-1, j-1) + P(x_i \diamond y_j) \\ A(i-1, j) \\ A(i, j-1) \end{cases}$$

but in general the solution is different from the Viterbi path.

Multiple alignment



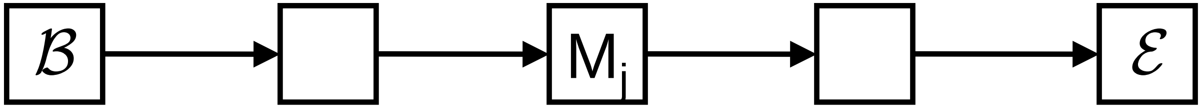
```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV
       ruler 1.......10........20........30........40........50........60........70........80........90
```

HMM for an aligned sequence family

- We regard the sequences as independent observations of a probabilistic model.
- Emission probabilities are different at each position of the alignment.
- The model defines a probability distribution over the whole sequence space.

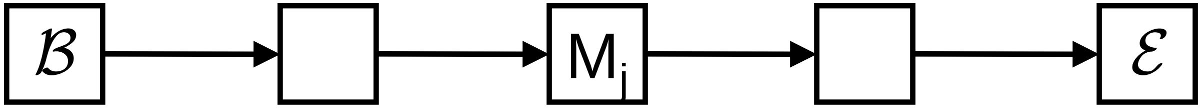
- Parameters:
  - Transition probabilities, *T*
  - Emission probabilities, *E*

Ungapped alignments



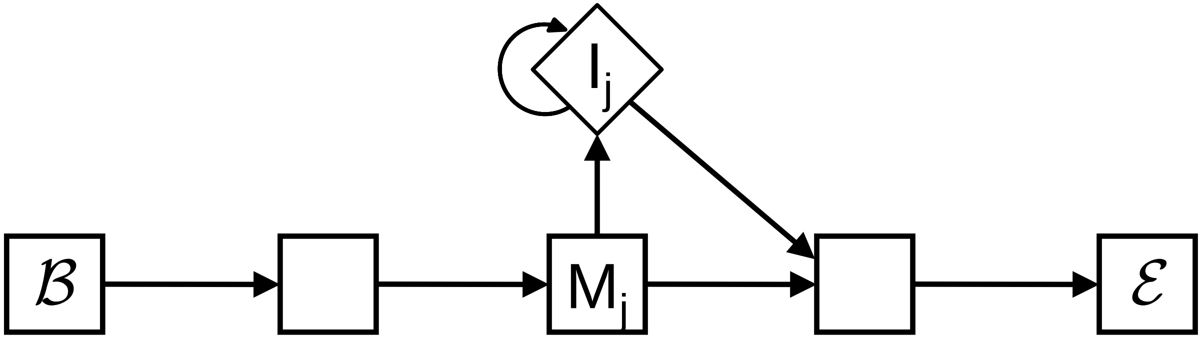
$$P(x \mid M) = \prod_{j=1}^{L} E_{M_j, x_j}$$

Log-odds score w.r.t. random model q



$$S(x) = \sum_{j=1}^{L} \log \frac{E_{M_j, x_j}}{q_{x_j}}$$

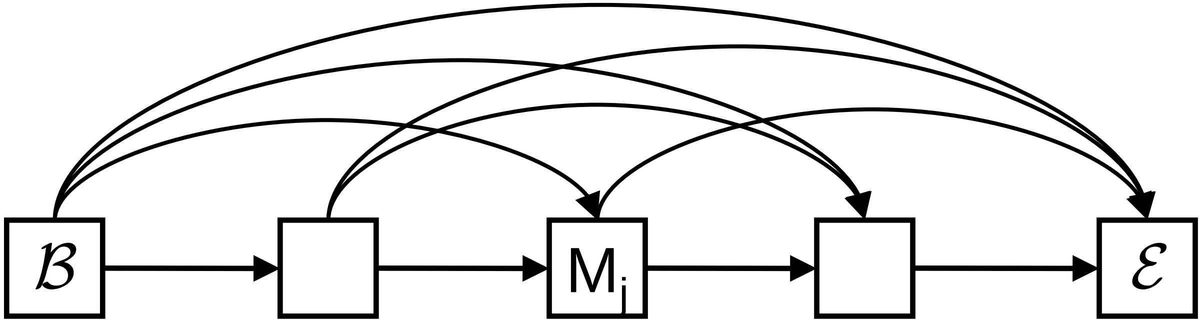
Insertions: affine gap scoring



- With $E_{I_j,a} = q_a$, emission probabilities cancel and the score of a gap of length k is

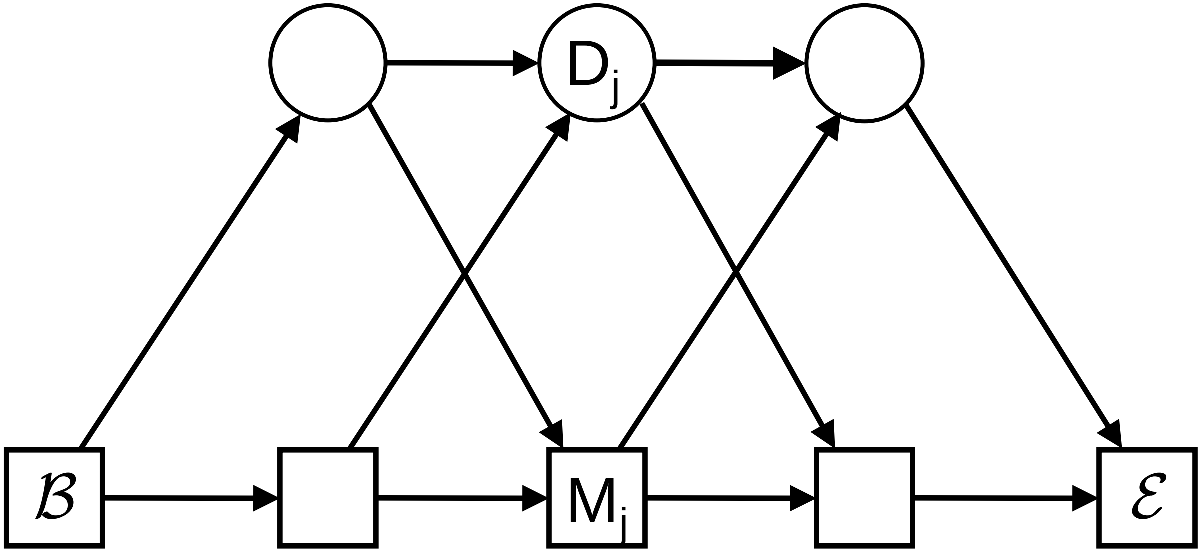
$$\log T_{M_j,I_j} + \log T_{I_j,M_{j+1}} + (k-1)\log T_{I_j,I_j}$$

Deletions



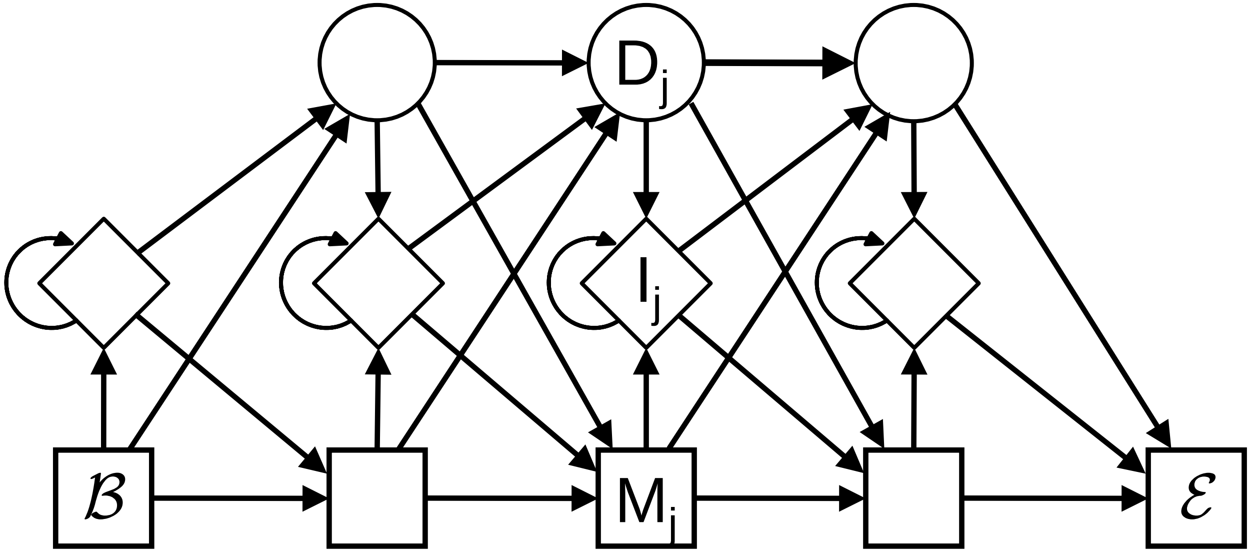
- This topology would require a lot of transitions

Profile HMM



- With silent delete states D_j, any jump can be realized by a series of $D_{j-1} \rightarrow D_j$ transitions.

The profile HMM



- The profile HMM is an unrolled version of the pair HMM.
- The profile HMM generalizes the pair HMM.
- I → D transitions are rare, but can be convenient to include.

Parameter estimation from a multiple alignment

- The parameters define a specific region of sequence space, for example, a protein family.
- Each protein family can be represented by a specific profile HMM (Pfam database, http://pfam.sanger.ac.uk/)
- The parameters of the profile HMM are
  - the length of the model, L
  - the transition probabilities, T
  - the emission probabilities, E

Length of the profile HMM

```
HBA_HUMAN  ...VGA--HAGEY...
HBB_HUMAN  ...V----NVDEV...
MYG_PHYCA  ...VEA--DVAGH...
GLB3_CHITP ...VKG------D...
GLB5_PETMA ...VYS--TYETS...
LGB2_LUPLU ...FNA--NIPKH...
GLB1_GLYDI ...IAGADNGAGV...
              ***  *****
```

- The length of the profile HMM corresponds to the expected number of Match states.
- Heuristic: count the number of columns with less than 50% gaps.

Transition and emission probabilities

```
HBA_HUMAN  ...VGA--HAGEY...
HBB_HUMAN  ...V----NVDEV...
MYG_PHYCA  ...VEA--DVAGH...
GLB3_CHITP ...VKG------D...
GLB5_PETMA ...VYS--TYETS...
LGB2_LUPLU ...FNA--NIPKH...
GLB1_GLYDI ...IAGADNGAGV...
              123  45678
```

- At each position, count the number of each transition, N_{kl}, and of each emission, N_{kx}. Then the ML estimates are

$$\hat{T}_{kl} = \frac{N_{kl}}{\sum_{l'} N_{kl'}} \quad \text{and} \quad \hat{E}_{kx} = \frac{N_{kx}}{\sum_{x'} N_{kx'}}$$

Membership detection

- Given
  - a profile HMM, $\mathcal{M}$, and
  - a new sequence x

  decide whether x belongs to the set of sequences (e.g., a protein family) represented by the HMM, or not.
- We can consider the most probable alignment

$$P(x, z^* \mid \mathcal{M})$$

or the full probability summing over all alignments

$$P(x \mid \mathcal{M}) = \sum_{z} P(x, z \mid \mathcal{M})$$

Log-odds scores

- Rather than the HMM probabilities, we consider the log-odds ratios with respect to the random model $\mathcal{R}$.

$$P(x \mid \mathcal{R}) = \prod_i q_{x_i}$$

- The random model assumes independent and identical character distributions q at each position i.
- We formulate Viterbi and Forward algorithms directly for the log-odds scores

$$\log \frac{P(x, z^* \mid \mathcal{M})}{P(x, z^* \mid \mathcal{R})} \quad \text{and} \quad \log \frac{P(x \mid \mathcal{M})}{P(x \mid \mathcal{R})}$$

Viterbi algorithm for profile HMMs

- Let $V^M(i,j)$ be the log-odds score of the best path for $x_{1\ldots i}$ ending in M_j and emitting x_i; and similarly for V^I and V^D. The Viterbi recursion is

$$V^M(i,j) = \log \frac{E_{M_j x_i}}{q_{x_i}} + \max \begin{cases} \log T_{M_{j-1},M_j} + V^M(i-1,j-1) \\ \log T_{I_{j-1},M_j} + V^I(i-1,j-1) \\ \log T_{D_{j-1},M_j} + V^D(i-1,j-1) \end{cases}$$

$$V^I(i,j) = \log \frac{E_{I_j x_i}}{q_{x_i}} + \max \begin{cases} \log T_{M_j,I_j} + V^M(i-1,j) \\ \log T_{I_j,I_j} + V^I(i-1,j) \\ \log T_{D_j,I_j} + V^D(i-1,j) \end{cases}$$

$$V^D(i,j) = \max \begin{cases} \log T_{M_{j-1},D_j} + V^M(i,j-1) \\ \log T_{I_{j-1},D_j} + V^I(i,j-1) \\ \log T_{D_{j-1},D_j} + V^D(i,j-1) \end{cases}$$

Viterbi algorithm for profile HMMs

- Simplifications:
  - $E_{I,x_i} = q_{x_i}$
  - no transitions D → I, I → D
- We allow the alignment to begin and to end in an Insert or Delete state.

- Initialization: $\mathcal{B} = M_0$ and $V^M(0,0) = 0$.
- Termination: $\mathcal{E} = M_{L+1}$ and

$$V^M(n, L+1) = \max \begin{cases} \log T_{M_L, M_{L+1}} + V^M(n-1, L) \\ \log T_{I_L, M_{L+1}} + V^I(n-1, L) \\ \log T_{D_L, M_{L+1}} + V^D(n-1, L) \end{cases} \qquad = \log \frac{P(x, z^* \mid \mathcal{M})}{P(x, z^* \mid \mathcal{R})}$$

Forward algorithm for profile HMMs

- Let $F^M(i,j)$ be the log-odds score of $x_{1\ldots i}$ ending in M_j and emitting x_i; and similarly for F^I and F^D. The Forward recursion is

$$
\begin{aligned}
F^M(i,j) &= \log \frac{E_{M_j,x_i}}{q_{x_i}} + \log \Big\{ T_{M_{j-1},M_j} \exp\left[F^M(i-1,j-1)\right] + \\
&\qquad + T_{I_{j-1},M_j} \exp\left[F^I(i-1,j-1)\right] + T_{D_{j-1},M_j} \exp\left[F^D(i-1,j-1)\right] \Big\} \\
F^I(i,j) &= \log \frac{E_{I_j,x_i}}{q_{x_i}} + \log \Big\{ T_{M_j,I_j} \exp\left[F^M(i-1,j)\right] + \\
&\qquad + T_{I_j,I_j} \exp\left[F^I(i-1,j)\right] + T_{D_j,I_j} \exp\left[F^D(i-1,j)\right] \Big\} \\
F^D(i,j) &= \log \Big\{ T_{M_{j-1},D_j} \exp\left[F^M(i,j-1)\right] + \\
&\qquad + T_{I_{j-1},D_j} \exp\left[F^I(i,j-1)\right] + T_{D_{j-1},D_j} \exp\left[F^D(i,j-1)\right] \Big\}
\end{aligned}
$$

- Note that, in general, $\log(p+q) = \log\left(e^{\log p} + e^{\log q}\right)$.

Multiple alignment with a known profile HMM

- Use Viterbi path to align each new sequence.
- The resulting multiple alignment separates characters emitted from Match and Insert states.
- Regions of Insert states correspond to poorly conserved or unalignable subsequences (e.g., coding for protein loops).
- Within Insert regions, characters are not aligned.

```
123    45678
VGAey..HAGEY
V--evd.NVDEV
VEAgh..DVAGH
VKGth..NV--D
VYSts..TYETS
FNAhk..NIPKH
IAGadgvNGAGV
```

unaligned

Multiple alignment from scratch

- Initialization:
  - Choose length of profile HMM
- Parameter estimation:
  - Use Baum-Welch algorithm to obtain MLEs of transition and emission probabilities:
    - E step: Run Forward and Backward algorithms to obtain expected transition and emission counts
    - M step: Estimate transition and emission probabilities
- Alignment:
  - Align all sequences to the model using the Viterbi algorithm
  - To build the alignment, remember, for each Insert state, the length of the longest inserted subsequence

Summary

- The pair HMM is the probabilistic graphical model for global (local) pairwise sequence alignment.
- The Viterbi algorithm corresponds to the Needleman-Wunsch (Smith-Waterman) algorithm.
- Profile HMMs are probabilistic graphical models of sequence families.
- Pair and profile HMM allow for reasoning probabilistically about sequence alignments. For example, we can ask for the probability of two characters being aligned, or for the probability of a given sequence being part of a known protein family, without relying on single optimal alignments.

References

- Durbin R, Eddy S, Krogh A, Mitchinson G. Biological Sequence Analysis. Cambridge University Press, 1998–2007. Chapters 4–6.
- Beerenwinkel N and Siebourg J. Statistics, probability, and computational science. In Maria Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, chapter 3, pages 77–110. Springer, New York, 2012. DOI: 10.1007/978-1-61779-582-4_3. Section 6.