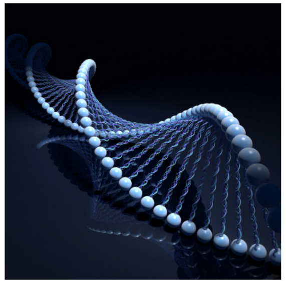
LGBIO2010: Multiple alignment - Profile HMMs

Pierre Dupont



UCL – ICTEAM

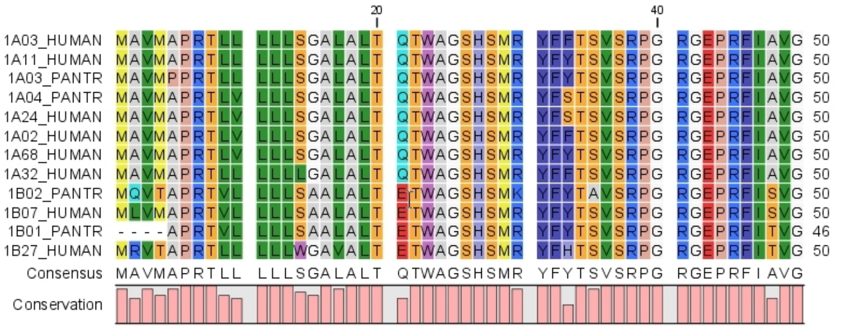
Outline

1. Multiple alignment
   - Computation of an optimal multiple alignment
   - Heuristic algorithms
2. Profile HMMs

Outline

1. Multiple alignment
   - Computation of an optimal multiple alignment
   - Heuristic algorithms
2. Profile HMMs

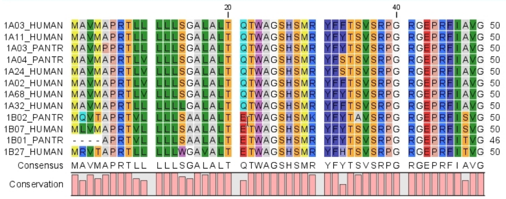
The Multiple Alignment Problem



```
                                       20                        40
                                        |                         |
  1A03_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
  1A11_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
   1A03_PANTR MAVMPPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
   1A04_PANTR MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
  1A24_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
  1A02_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
  1A68_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
  1A32_HUMAN  MAVMAPRTLL  LLLLGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
   1B02_PANTR MQVTAPRTVL  LLLSAALALT  ETWAGSHSMK  YFYTAVSRPG  RGEPRFISVG  50
  1B07_HUMAN  MLVMAPRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFISVG  50
   1B01_PANTR ----APRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFITVG  46
  1B27_HUMAN  MRVTAPRTLL  LLLWGAVALT  ETWAGSHSMR  YFHTSVSRPG  RGEPRFITVG  50
   Consensus  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG
Conservation
```

- align 3 or more homologous sequences
- either globally or locally (look only for conserved segments)

Scoring a multiple alignment

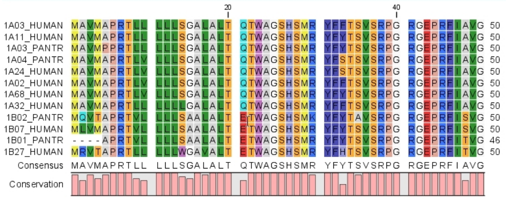


```
                                        20                  40
                                         |                   |
1A03_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A11_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A03_PANTR  MAVMPPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A04_PANTR  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A24_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A02_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A68_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A32_HUMAN  MAVMAPRTLL  LLLLGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1B02_PANTR  MQVTAPRTVL  LLLSAALALT  ETWAGSHSMK  YFYTAVSRPG  RGEPRFISVG  50
1B07_HUMAN  MLVMAPRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFISVG  50
1B01_PANTR  ----APRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFITVG  46
1B27_HUMAN  MRVTAPRTLL  LLLWGAVALT  ETWAGSHSMR  YFHTSVSRPG  RGEPRFITVG  50
 Consensus  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG
Conservation
```

Assumption

- Individual columns are assumed statistically independent
- A multiple alignment m with L columns can then be scored as
$$S(m) = G + \sum_{i=1}^{L} S(m_i)$$
  - $S(m_i)$ = score for column i
  - G = score for all gaps in m using linear or affine gap penalties

SP score



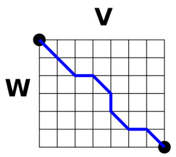
```
                                        20                  40
                                         |                   |
1A03_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A11_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A03_PANTR  MAVMPPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A04_PANTR  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A24_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A02_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A68_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A32_HUMAN  MAVMAPRTLL  LLLLGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1B02_PANTR  MQVTAPRTVL  LLLSAALALT  ETWAGSHSMK  YFYTAVSRPG  RGEPRFISVG  50
1B07_HUMAN  MLVMAPRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFISVG  50
1B01_PANTR  ----APRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFITVG  46
1B27_HUMAN  MRVTAPRTLL  LLLWGAVALT  ETWAGSHSMR  YFHTSVSRPG  RGEPRFITVG  50
 Consensus  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG
Conservation
```

Sum of pairs

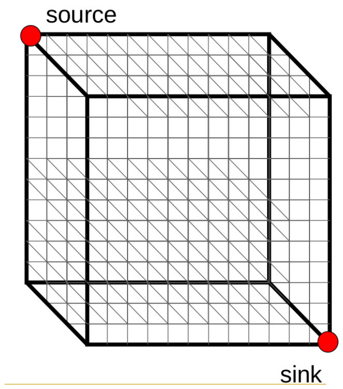
- Column score:
$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$
where $s(a,b)$ is given by a substitution scoring matrix (*e.g.* PAM or BLOSUM)
- Gap penalty
  - linear: $s(a,-) = s(-,b) = -d$; $s(-,-) = 0$
  - affine: all gaps are scored separately

Optimal alignment through dynamic programming

From 2 to 3 sequences



2-D edit lattice

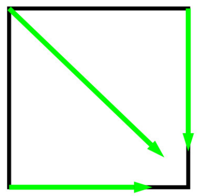


3-D edit lattice

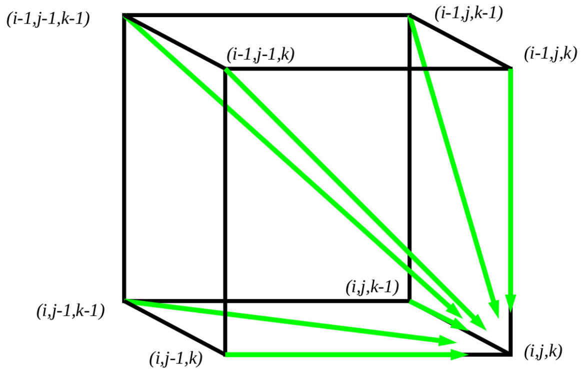
Illustrations from www.bioalgorithms.info

Optimal alignment through dynamic programming

From 2 to 3 sequences



$2^2 - 1 = 3$ possible moves



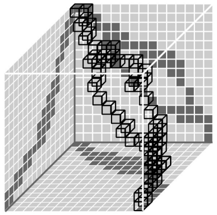
$2^3 - 1 = 7$ possible moves

Illustrations from www.bioalgorithms.info

Dynamic programming

- Optimal alignment between k sequences can be computed through dynamic programming
- The time complexity for k sequences of average length $\bar{n}$ is in $O(2^k \bar{n}^k)$ and the space complexity in $O(\bar{n}^k)$ (for storing the hyper-cube)
  - in practice, computation must be limited to very few sequences due to the exponential growth with k

MSA algorithm



- MSA is an optimized DP algorithm which first computes all pairwise alignments and then limits the exploration of the (hyper-)cube to regions consistent with those alignments
  - time complexity in $O(k^2 \bar{n}^2)$ but somewhat complex to program
  - MSA can optimally align $\approx$ 10 sequences of up to 200-300 residues in *reasonable time*
- a recent parallel extension G-MSA is reported to align up to 500 sequences of 236 residues on average within 10 seconds on a Linux machine including 2 cores with GPUs [J. Blazewicz et al., 13]

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

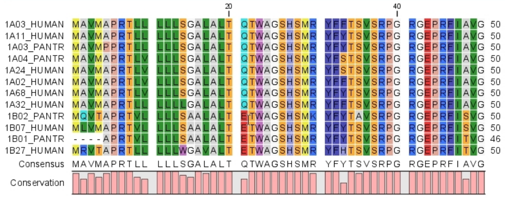
Progressive alignment methods

Greedy heuristic algorithms

succession of pairwise alignments

- 2 sequences are aligned first
- a sequence is added to a group of already aligned sequences
  - compute all pairwise alignments between s and an existing group g of aligned sequences
  - the highest scoring pairwise alignment determines how the new sequence s is aligned to the group g
- a group g_1 of sequences is aligned to another group g_2 of sequences
  - all sequence pairs between g_1 and g_2 are tried
  - the best pairwise alignment determines the alignment of both groups

Issues with progressive pairwise alignments



```
                                             20                                40
                                             |                                 |
  1A03_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
  1A11_HUMAN  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
  1A03_PANTR  MAVMPPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
  1A04_PANTR  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
  1A24_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
  1A02_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
  1A68_HUMAN  MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
  1A32_HUMAN  MAVMAPRTLL  LLLLGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
  1B02_PANTR  MQVTAPRTVL  LLLSAALALT  ETWAGSHSMK  YFYTAVSRPG  RGEPRFISVG  50
  1B07_HUMAN  MLVMAPRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFISVG  50
  1B01_PANTR  ----APRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFITVG  46
  1B27_HUMAN  MRVTAPRTLL  LLLWGAVALT  ETWAGSHSMR  YFHTSVSRPG  RGEPRFITVG  50
   Consensus  MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG
Conservation
```

When aligning a new sequence to an existing group

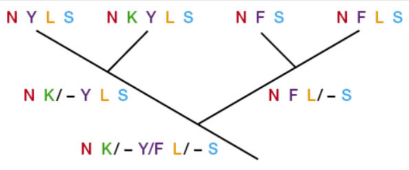
1. the degree of sequence conservation at each position should be taken into account
2. mismatches at highly conserved positions should be more penalized
3. the order in which sequences are incorporated in the multiple alignment matters

Those aspects are ignored by the sum of pairs scoring

ClustalW

Main steps

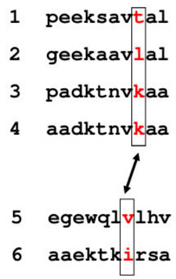
1. construct a distance matrix of all $\frac{k(k-1)}{2}$ pairwise alignment scores
   - correct those scores by considering the Kimura evolutionary model (see phylogeny)
2. build a tree using the neighbor-joining algorithm (see phylogeny)



3. use it as a guide tree: progressively align nodes of decreasing similarity
   - sequence-sequence, sequence-profile and profile-profile alignments

Illustration from Bioinformatics: Sequence and Genome Analysis, 2nd edition (© Cold Spring Harbor Lab. Press 2004)

ClustalW: weighted Sum-of-Pairs score



```
1  peeksavtal
2  geekaavlal
3  padktnvkaa
4  aadktnvkaa

5  egewqlvlhv
6  aaektkirsa
```

$$\begin{aligned} s(m_i) &= s(t,v) \\ &+ s(t,i) \\ &+ s(l,v) \\ &+ s(l,i) \\ &+ s(k,v) \\ &+ s(k,i) \\ &+ s(k,v) \\ &+ s(k,i) \end{aligned}$$

$$\begin{aligned} s(m_i) &= s(t,v) * w_1 * w_5 \\ &+ s(t,i) * w_1 * w_6 \\ &+ s(l,v) * w_2 * w_5 \\ &+ s(l,i) * w_2 * w_6 \\ &+ s(k,v) * w_3 * w_5 \\ &+ s(k,i) * w_3 * w_6 \\ &+ s(k,v) * w_4 * w_5 \\ &+ s(k,i) * w_4 * w_6 \end{aligned}$$

Weights are derived from the guide tree:
the more distant the sequences the higher the weighting

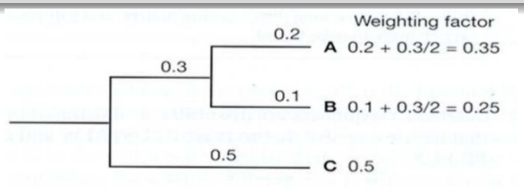


Illustration from Bioinformatics: Sequence and Genome Analysis, 2nd edition (© Cold Spring Harbor Lab. Press 2004)

ClustalW

Further heuristics

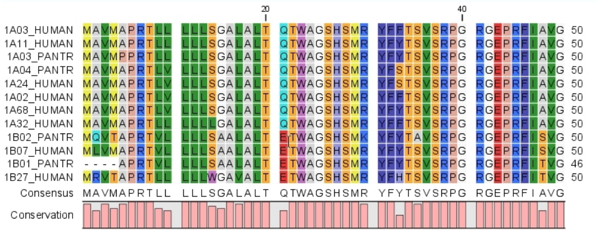
- Position-specific gap-open penalties with decreased penalties wherever other gaps have already been found among already aligned sequences
- gap penalties are also decreased or increased based on a large collection of structural alignments

  - as a special case, hydrophobic residues (more likely to be buried) are associated to higher gap penalties
- the guide tree may be adjusted on the fly to defer a low scoring alignment until more profile information has been accumulated

Outline

1. Multiple alignment
2. Profile HMMs

Beyond multiple alignments



```
                               20                                    40
                                |                                     |
1A03_HUMAN   MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A11_HUMAN   MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A03_PANTR   MAVMPPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A04_PANTR   MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A24_HUMAN   MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFSTSVSRPG  RGEPRFIAVG  50
1A02_HUMAN   MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1A68_HUMAN   MAVMAPRTLV  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG  50
1A32_HUMAN   MAVMAPRTLL  LLLLGALALT  QTWAGSHSMR  YFFTSVSRPG  RGEPRFIAVG  50
1B02_PANTR   MQVTAPRTVL  LLLSAALALT  ETWAGSHSMK  YFYTAVSRPG  RGEPRFISVG  50
1B07_HUMAN   MLVMAPRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFISVG  50
1B01_PANTR   ----APRTVL  LLLSAALALT  ETWAGSHSMR  YFYTSVSRPG  RGEPRFITVG  46
1B27_HUMAN   MRVTAPRTLL  LLLWGAVALT  ETWAGSHSMR  YFHTSVSRPG  RGEPRFITVG  50
 Consensus   MAVMAPRTLL  LLLSGALALT  QTWAGSHSMR  YFYTSVSRPG  RGEPRFIAVG
```

Motivations

- multiple alignments are most often based on pairwise alignments
- a new sequence x may be only distantly and locally related to each sequence in a known family (biological question)
  - all pairwise alignments between x and each family members may look poor
  - need to model statistical features shared by the family members
- computing the alignment between x and a probabilistic model of the family may be much more efficient computationally

Probabilistic models of a known family

Questions

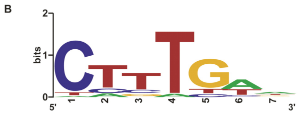
1. given a multiple alignment between sequences how to build a global/local model M from it?
2. how to compute the matching score between a query sequence x and the model M?
3. how to get rid of the initial alignment?

Position-specific scoring matrices (PSSM)

A

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

B



- Local model for a window length L and ungapped score matrix from N sequences

$$P(x|M) = \prod_{i=1}^{L} P(x_i|M) = \prod_{i=1}^{L} \frac{f(x_i)}{N}$$

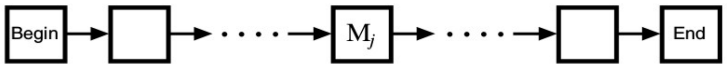
- Log-odds score

$$S = \sum_{i=1}^{L} \log \frac{P(x_i|M)}{q_{x_i}}$$

  - q_{x_i} = the background model (*e.g.* multinomial model)
- One evaluates the score S between x and M for all positions x_i and a sliding window of size L

Illustration from http://sites.google.com/site/iiserbioinformatics/tutorials

PSSMs are very simple HMMs

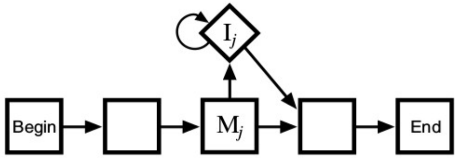


$$P(x|M) = \prod_{i=1}^{L} P(x_i|M)$$

- $P(x_i|M)$ are emission probabilities on match states
- transition probabilities are all equal to 1 (linear structure)
- need to account for possible gaps
- better to avoid a prescribed window length L

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

Adding insert states

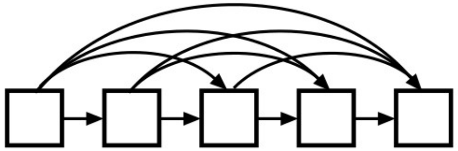


- to account for portions of x that do not match the model
- emission probabilities on insert states are typically defined through the background model
  - no contribution to the log-odds score $\log \frac{P(x_i|I_j)}{q_{x_i}} = \log \frac{q_{x_i}}{q_{x_i}} = 0$
- transition probabilities to insert states and back are equivalent to affine gap penalties $\log \mathbf{A}_{M_j I_j} + \log \mathbf{A}_{I_j M_{j+1}} + (k-1) \log \mathbf{A}_{I_j I_j}$

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

Adding delete states

- portions of the model M that are not matched by any residue x_i could be modeled by skipping transitions



- to allow arbitrary long gaps it is more convenient to introduce delete states which are silent states

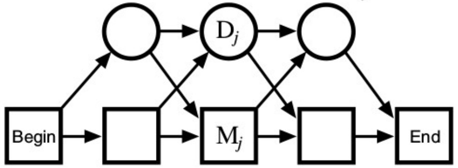
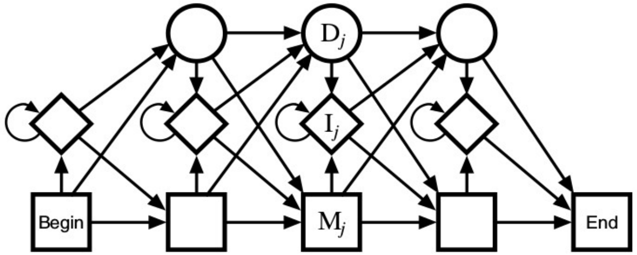
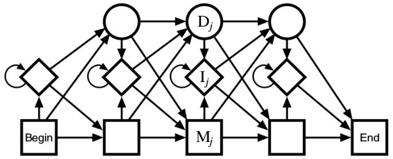


Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

A full profile HMM



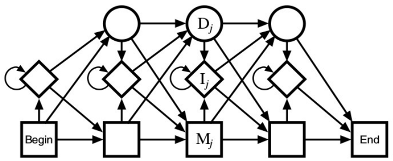
Deriving a pHMM from a multiple alignment



```
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...
               ***  *****
```

- a match state for each conserved position (*e.g.* at least 50%)
- insert states: *e.g.* columns with at least 50% gaps
- delete states: gaps on match positions
- emission (for match states) and transition probabilities are estimated from the counts
  - $\mathbf{B}_{ki} = \frac{f(k,i)}{f(k)}$
    - $f(k,i)$ = number of times symbol i is observed on state k
    - $f(k)$ = number of times state k is used
  - $\mathbf{A}_{kl} = \frac{f(k,l)}{f(k)}$
    - $f(k,l)$ = number of times a transition from state k to state l is used

Smoothing probability estimates



```
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...
                ***  *****
```

Whenever the initial multiple alignment is limited to a few sequences, some emission/transition probabilities may be null

Additive smoothing with pseudo-counts

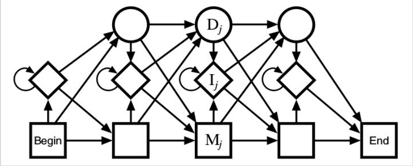
- $\mathbf{B}_{ki} = \frac{f(k,i)+\varepsilon}{f(k)+\sum_i \varepsilon}$ with *e.g.* $10^{-6} \leq \varepsilon \leq 1$
  - $f(k,i)$ = number of times symbol i is observed on state k
  - $f(k)$ = number of times state k is used
- $\mathbf{A}_{kl} = \frac{f(k,l)+\varepsilon'}{f(k)+\sum_l \varepsilon'}$ with *e.g.* $10^{-6} \leq \varepsilon' \leq 1$
  - $f(k,l)$ = number of times a transition from state k to state l is used

Unsupervised learning

Objective: no need for an initial multiple alignment but just a collection of unaligned sequences

Procedure

1. choose a general pHMM structure

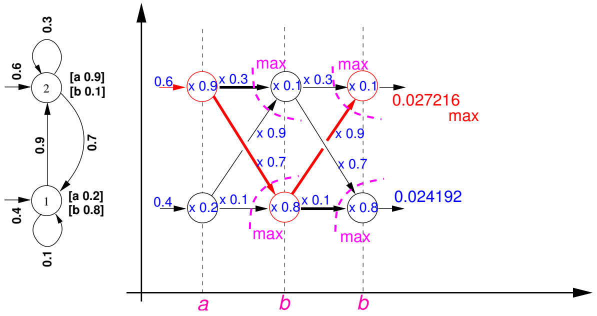


2. choose the number of match states:
   e.g. half the average sequence length
3. estimate the pHMM parameters through Viterbi or Baum-Welch

Matching a sequence to a pHMM

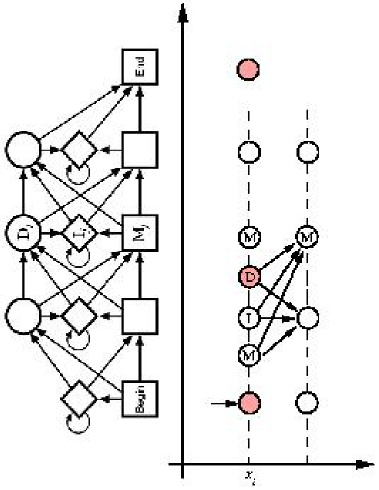
Viterbi recurrence

- Computations are done usually with log's:
  $$-\log \gamma(k,t) = \min_l[-\log \gamma(l,t-1) - \log \mathbf{A}_{lk}] - \log \mathbf{B}_{kx_t}$$
- Including a background model to produce a log-odds score:
  $$\log \gamma(k,t) = \max_l[\log \gamma(l,t-1) + \log \mathbf{A}_{lk}] + \log \frac{\mathbf{B}_{kx_t}}{q_{x_t}}$$
- A similar adaptation can be included into the forward recurrence

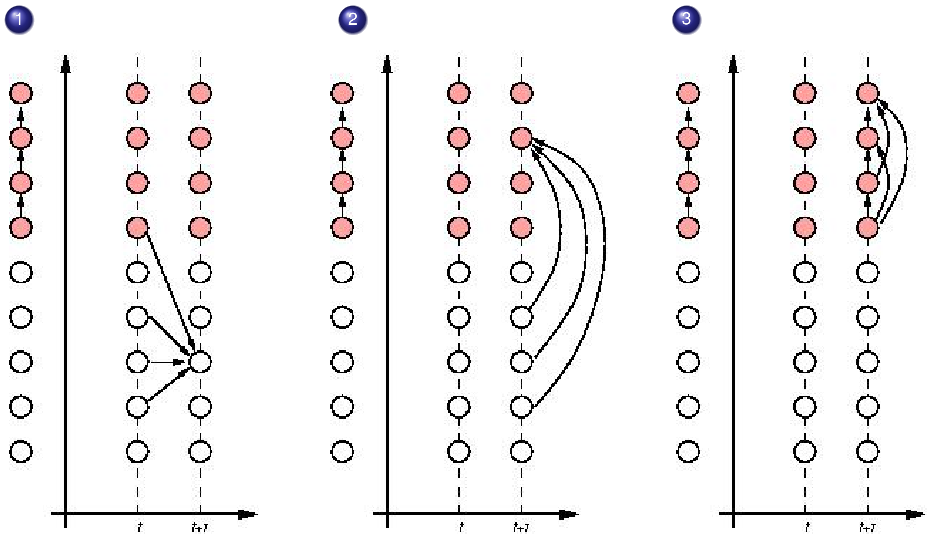


Silent states

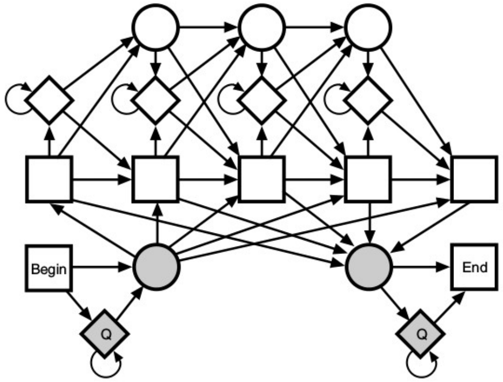
Begin, End and D states are silent = non-emitting



Computing with silent states



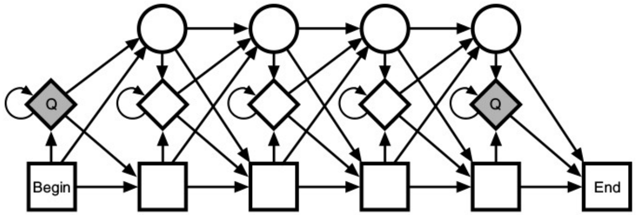
pHMM for non-global alignments



- non-conserved fragments are modeled through flanking insert states using the background emission probabilities
- flanking delete states allow for starting or ending the profile at any point

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

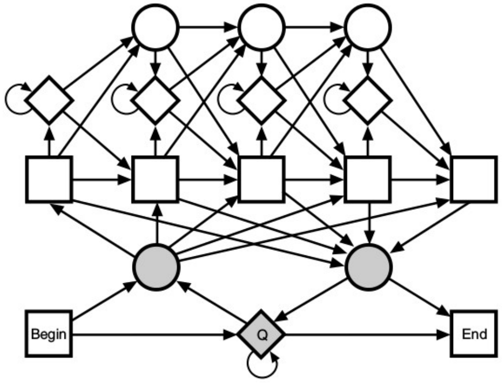
pHMM for non-global alignments



- forcing the match of the complete profile (or own delete states)
- no flanking delete states

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

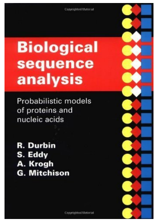
pHMM for non-global alignments



- allowing repeated matches to subsections of the profile

Illustration from Biological Sequence Analysis (© Cambridge University Press 1998)

Further reading



Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

- Chapter 5: Profile HMMs for sequence families
- Chapter 6: Multiple sequence alignment methods

Blazewicz, J., Frohmberg, W., Kierzynka, M. and Wojciechowski, P.
G-MSA - A GPU-based, fast and accurate algorithm for multiple sequence alignment
Journal of Parallel Distributed Computing, Vol. 73, p. 32–41, (2013).
http://dx.doi.org/10.1016/j.jpdc.2012.04.004

Database and software tools

- Multiple Sequence Alignment by CLUSTALW
  http://www.genome.jp/tools/clustalw/
- PFAM: database a protein families represented as MSA and HMMs
  http://pfam.xfam.org/
- HMMER: biosequence analysis with profile HMMs
  http://hmmer.org/