SGN-6156, Lecture 4
Biological sequence analysis

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

Department of Signal Processing,

Tampere University of Technology

09.04.2008

The probability of a sequence

- Another useful quantity is the probability of a sequence $P(x)$ (i.e., given a HMM, $P(x|\theta)$)
- For example, that allows (among many other things) to compare different HMMs using Bayesian model comparison
- A sequence of symbols x can be generated via several paths, thus

$$P(x) = \sum_{\pi} P(\pi, x)$$

- Let $f_k(i)$ denote the probability of the observed subsequence $(x_1, \ldots, x_i)$ such that $\pi_i = k$, i.e.,

$$f_k(i) = P(x_1, \ldots, x_i, \pi_i = k)$$

- The probability of $f_l(i+1)$ for all l can be found as

$$f_l(i+1) = \left[\sum_k f_k(i) a_{kl} \right] e_l(x_{i+1})$$

The forward algorithm

- Initialization: $i = 0$, $f_k(0) = 0$ for $k > 0$
- Recursion: $i = 1, \ldots, L$, for all l

$$f_l(i) = \left[\sum_k f_k(i-1) a_{kl} \right] e_l(x_i)$$

- Termination:

$$P(x) = \sum_k f_k(L) a_{k0}$$

The probability of a state

- Yet another interesting quantity is the probability that observation x_i is emitted from state k, i.e., $P(\pi_i = k|x)$
- First compute the probability of $(\pi_i = k, x)$

$$\begin{aligned} P(\pi_i = k, x) &= P(x_1, \ldots, x_i, \pi_i = k) P(x_{i+1}, \ldots, x_L | x_1, \ldots, x_i, \pi_i = k) \\ &= P(x_1, \ldots, x_i, \pi_i = k) P(x_{i+1}, \ldots, x_L | \pi_i = k) \\ &= f_k(i) b_k(i) \end{aligned}$$

- $f_k(i)$ is the quantity used in the forward algorithm
- $b_k(i)$ can be computed similarly, so called backward algorithm
- From the definition of conditional probability one gets

$$P(\pi_i = k|x) \frac{P(\pi_i = k, x)}{P(x)} = \frac{f_k(i) b_k(i)}{P(x)}$$

The backward algorithm

- Initialization: $i = L$, $b_k(L) = 0$ for all k
- Recursion: $i = L, \ldots, 1$, for all k

$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$

- Termination:

$$P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$$

- See Figures 3.6 and 3.7 in (Durbin et al., 1998)

Posterior decoding

- Instead of the Viterbi solution π^*, one can use the

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k|x)$$

- This can be more appropriate than π^* if there are several paths that have approximately the same probability
- Note that $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_L)$ may even represent an impossible path, i.e., $P(\hat{\pi}|x) = 0$

Parameter estimation for HMMs

- HMMs contains transition and emission probabilities, a_{kl} and $e_k(b)$
- Parameters can be estimated from data (both supervised and unsupervised)
- We will skip this interesting and important topic for now but will get back to this topic later on if needed...

Choice of HMM model structure

- All previous model structures have been fully connected
- In applications, HMM model structure is typically constructed by hand
- If e.g. transitions from state k to state l are not allowed, then simply set $a_{kl} = 0$
- Some model structures are shown on page 69 in (Durbin et al., 1998)
- The HMM model structure can also be learned from training data as well
- Let M_i denote the HMM structure and θ_i its parameters
- A simple approach: if there is lots of data, then compute $P(x|M_i, \theta_i)$ and consequently e.g.

$$P(M_i, \theta_i|x) = \frac{P(x|M_i, \theta_i)P(M_i, \theta_i)}{\sum_i P(x|M_i, \theta_i)P(M_i, \theta_i)}$$

Silent states

- States that do not emit symbols
- These can be useful for reducing the complexity of the model
- See an example on pages 70–71 in (Durbin et al., 1998)

Numerical stability of HMMs

- Long sequences would require extremely high numerical precision
- Two general techniques to avoid numerical instability
  - The $\log$-transformation
  - Scaling of probabilities

Pairwise alignment using HMMs

- The material below is mainly based on Section 4 in (Durbin et al., 1998)
- In the case of the affine gap penalty, we used finite state machines (FSA) to align a sequence pair
- FSAs can be converted into HMMs relatively easily
- HMMs provide truly probabilistic interpretation of pairwise alignments allowing assessment of
  - Reliability of alignments
  - Sample alternative suboptimal alignments

- Convert a FSA to a HMM by
  - Assigning probabilities to transitions between states and emission of symbols from states
  - Define start and end states
- See Figures 4.1–4.2 in (Durbin et al., 1998)
- This is similar with HMMs introduced before except that instead of emitting a sequence x this pair HMM generates a pairwise alignment
- The standard HMM algorithms can be applied with an extra dimension (e.g. $v_k(i, j)$ instead of $v_k(i)$)

The most probable alignment

- Viterbi algorithm can again be applied to find the most probably path which corresponds to the optimal FSA alignment
- As above, $v^{\bullet}(i,j)$ denotes the probability of the most probably path ending in $\bullet$ and emitting symbols x_i and y_j

Viterbi for pair HMMs

- Initialization: $v^M(0,0) = 1$, and $v^\bullet(i,0) = v^\bullet(0,j) = 0$ for all i, j, and $\bullet \in \{M, X, Y\}$
- Recursion: $i = 1, \ldots, n$, $j = 1, \ldots, m$

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

- Termination:

$$v^E = \tau \max(v^M(n,m), v^X(n,m), v^Y(n,m))$$

- Optimal path/alignment can be found by keeping track of pointers and backtracking
- A related HMM can also be constructed for
  - The random model (i.e., for unrelated sequences)
  - Local alignment (see Figure 4.3 in (Durbin et al., 1998)
  - etc.

The probability of aligning x and y

- If there is just one high-scoring alignment, then the best alignment is representative and the score itself useful
- When x and y are not closely related, then choosing a low-scoring alignment can be misleading, see Figure 4.4 in (Durbin et al., 1998) (this is a useful guideline even more generally)
- HMM framework provides a way to compute the probability of any alignment π

$$P(x,y) = \sum_{\pi} P(\pi, x, y)$$

- As in the case of standard HMMs, we can use the forward algorithm to compute $P(x,y)$ efficiently
- Let $f^k(i,j)$ denote the probability of all possible alignments up to (i,j) that end with state k

Forward algorithm for pair HMMs

- Initialization: $f^M(0,0) = 1$, $f^X(0,0) = f^Y(0,0) = 0$ and all $f^\bullet(i,-1) = f^\bullet(-1,j) = 0$
- Recursion: $i = 0,\ldots,n$, $j = 0,\ldots,m$ except $(0,0)$

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

- Termination:

$$f^E = \tau[f^M(n,m) + f^X(n,m) + f^Y(n,m)]$$

Posterior distribution of alignments

- The probability of any alignment can be used to compute the posterior distribution of an alignment as

$$P(\pi|x,y) = \frac{P(\pi,x,y)}{P(x,y)}$$

- As mentioned above, $P(\hat{\pi}|x,y)$ can be remarkably small
- If x and y are unrelated, then a low probability is of course understandable (even desired)
- A low value of $P(\hat{\pi}|x,y)$ can be due to the fact that there are many small variants π that have almost the same probability

Suboptimal alignment

- Find particular alignments whose probability is close to the most probable one
- Two different types of suboptimal alignments
  - Alignments differ in only a few positions
  - A major difference
- A general strategy: sample alignments from the posterior
- Sampling performed when tracing back $f^M(i,j)$
- Sampled alignments $\pi_1, \pi_2, \ldots$ can be used to estimate any interesting quantity
- Another method can be used to find distinct alignments. The method works by repeatedly modifying the Viterbi matrix and setting the score of previously sampled paths/alignments to zero.

References

- R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.