Hidden Markov Models and BioInformatics, Part II

Steven R. Dunbar

March 16, 2017

Outline

Review: dishonest casino as HMM

- the casino uses a fair die most of the time,
- occasionally the casino secretly switches to a loaded die,
- later the casino switches back to the fair die.
- the switch from fair-to-loaded occurs with probability 0.01
- from loaded-to-fair with probability 0.02.
- assume that the loaded die will come up "six" with probability 0.5
- the remaining five numbers with probability 0.1 each.

Fundamental Biological Problems

1. Infer the functions and structure of proteins based on their amino acid (residue) sequences.
2. Group proteins into families with similar functions and structure.
3. Construct phylogenetic trees for proteins showing inferred evolutionary relationships.

Differences from last week

1. Proteins, not DNA
2. 20 amino acids (also called residues), not 4
3. More proteins, but they are shorter (approximately 30,000 to 40,000 tabulated human proteins, average length is about 400, length range is 100 to 2000.)

Protein Similarity

For a pair (or more) of proteins, an important question is: *How are the proteins similar?*

1. Detect and measure overall similarity between protein amino acid sequences.
2. Find proteins with similar functions in different organisms by very similar subsequences of amino acids, called "conserved sequences".
3. Detect conserved sequences and evolution of conserved sequences.

Alignment is the method for answering these questions.

Simplest Alignment Problem

$x = x_1 \dots x_k$ from finite alphabet N

$y = y_1 \dots y_l$ from finite alphabet $M \supset N$

$k \ll l$

Find x in y.

This is the *text editor string search problem*:

Scan y for x_1, then carry on

Simplest Gap Alignment (Wildcards)

Introduce wild card entries $x_1 \dots x_s * x_{s+1} \dots x_k$

First search for indices j_1 with $y_{j_1+i} = x_i$, then $j_2 > j_1$ with $y_{j_2+i} = x_{s+i}$

Simplifying assumptions:

1. Where the wild-card gaps are is specified in advance
2. Insist on perfect matches

Alignment for DNA and Proteins

There are two types of alignment:

- A global alignment is an alignment of the full length of two sequences.
- A local alignment is an alignment of part of one sequence to part of another sequence.
- For (possibly) distantly related sequences, it might be more sensible to make local alignments of subregions of high similarity, not the whole sequence
- Allow introduction of gaps

Optimal Gapped Alignment 1

$$x = x_1 \dots x_k \text{ from finite alphabet } N$$
$$y = y_1 \dots y_l \text{ from finite alphabet } M \supset N$$
$$k \leq l$$
$$N = M = \{A, C, T, G\}$$
$$x = ACACTGT,$$
$$y = TAGACGGAGCTTCAC$$

Find "best" match of x with y.

Optimal Gapped Alignment 2


```
      A  C  -  -  A  C  -  T  G  T
T  A  G  A  C  G  G  A  G  C  T  -  T  C  A
```

- Introduce gaps (if necessary) within both sequences
- Allow matches and mismatches (with various scores based on chemistry and biology)
- Penalize (somewhat) for introducing gaps

Toy Example

Align `GAATTC` with `GATTA`, allowing gaps.

Score $+2$ for match, -1 for mismatch, -2 to gap.

G A A T T C

G A T T - A

Score of $2 + 2 - 1 + 2 - 2 - 1 = 2$

G A A T T C

G A - T T A

Score of $2 + 2 - 2 + 2 + 2 - 1 = 5$

Algorithms for Alignment

- Needleman-Wunsch for optimal global alignment, uses dynamic programming.
- Smith-Waterman for optimal local alignment, uses dynamic programming.
- N-W time is quadratic in length, S-W time is cubic in length, hence unsuitable for long sequences

Statistical Similarity 1

Analogous problem: Given two sequences of H and T, did the same coin produce both sequences?

Certainly can't match H to H, etc., but the statistical properties of each sequence can help accept or reject the possibility that same coin produced both sequences.

What statistics can we create for protein sequences? HMMs may be appropriate since they produce likelihoods of sequences.

Statistical Similarity 2

The Needleman-Wunsch alignment algorithm will produce a global alignment even if we give it two very distantly related (or unrelated) protein sequences, although the alignment score would be low.

But is this alignment statistically significant? In other words, is this alignment better than we would expect between any two random proteins?

Multiple Alignment

A common task in bioinformatics is to obtain a cluster of related sequences (e.g. from a database), and then to align those sequences using multiple alignment algorithms.

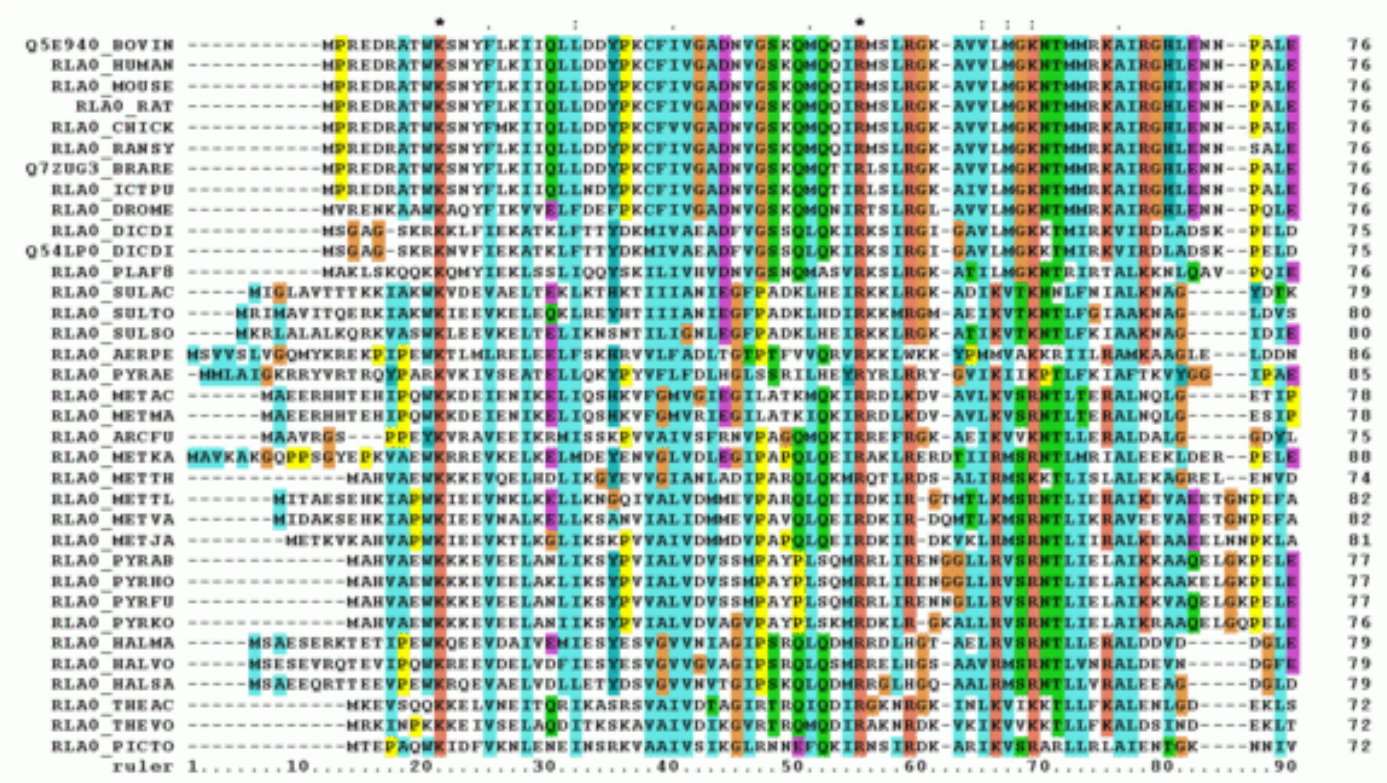
The clustering reflects the insights of the biology community as to which proteins belong within the same family. The outcome of the clustering process is a set of distinct protein families.

This is the first step in most phylogenetic analyses.

Multiple alignment time increases exponentially with the number of sequences.

Multiple Alignment

Alignment of acidic ribosomal protein P0 from several organisms.



Multiple Alignment Algorithms and Databases

Heuristic algorithms are generally used.

- CLUSTAL family of algorithms
- COFFEE family
- MUSCLE family
- MAFFT

There are large databases of proteins and alignments. (Some created with HMMs, some provide HMM data, see below.)

Definition of Profile

A profile HMM (pHMM) is a particular Hidden Markov Model (states, signals, transition matrix, and emission matrix) summarizing a multiple sequence alignment.

```
VGA--HAGEY
V----NVDEV
VEA--DVAGH
VKG------D
VYS--TYETS
FNA--NIPKH
IAGADNGAGV
```

Structure of a Profile 1

Profile HMMs have three states for each alignment position (i.e. each column in the MSA)

Model three possible outcomes when aligning each residue of the query sequence with the MSA.

- The query residue may align (match) with the next residue of the MSA,
- it may correspond to an insertion (new residue) relative to the MSA,
- it may correspond to a deletion (a gap) relative to MSA.

Structure of a Profile 2

Heuristic rule assigning MSA columns as match states: match column if less than half of the characters are gaps. Using this heuristic columns 1-3 and 6-10 are match columns. Length of pHMM is number of columns in the MSA assigned to match states, so length of the pHMM is 8.

```
VGA--HAGEY
V----NVDEV
VEA--DVAGH
VKG------D
VYS--TYETS
FNA--NIPKH
IAGADNGAGV
```

Match State Emissions in a pHMM

The most basic state is the match state, which matches (i.e. aligns) query residues at a specific position (column) in the MSA.

Each match state in the pHMM has its own corresponding set of emission probabilities, generated from counting the frequencies of each amino acid in the corresponding column.

Insertion States in a pHMM

For insertions, i.e. portions of the query sequence that do not match anything in the multiple alignment, an insert state is added.

As in the case of the match states, each insert state has its own set of emission probabilities. The insert state emission probabilities are typically generated using the distribution of amino acids over the entire MSA.

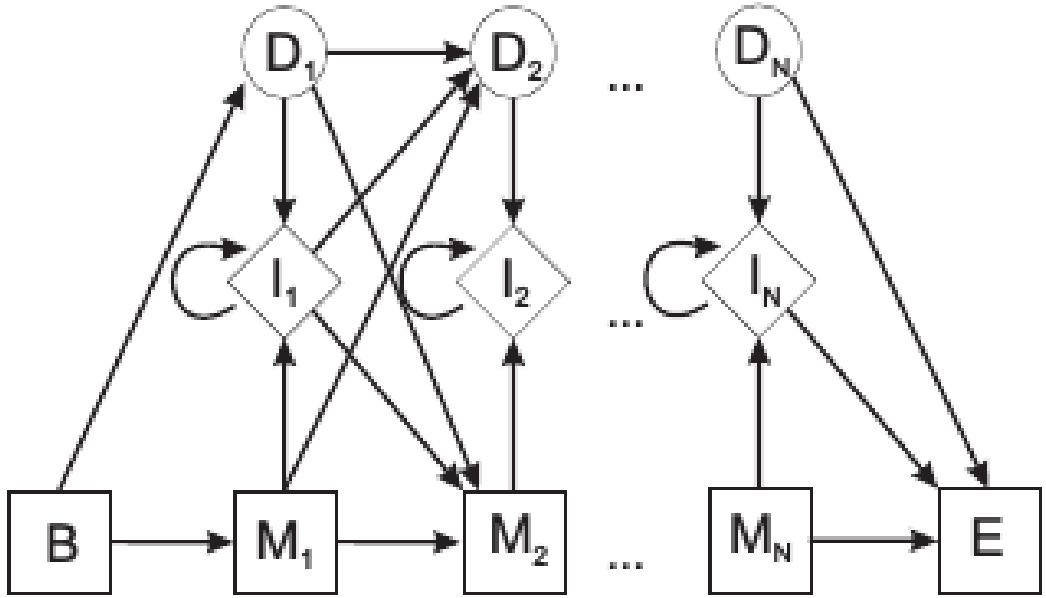
Delete States in a pHMM

A delete state is possible for each of the positions in the MSA.

The delete state is an example of a silent state in the model, as it does not emit any residues.

Structure of a Profile 3

Let l denote the number of match locations. Then the associated profile HMM has $3l + 3$ states in the underlying Markov process, namely: A "start" state S , an "end" state E , l match states $M_1, \ldots, M_l$, l "delete" states $D_1, \ldots, D_l$, and $l + 1$ "insert" states $I_0, \ldots I_l$.



Application of a pHMM

Start with collection of protein families (clusters) $F_1 \ldots F_k$, where all proteins within a family have the same length (after assigning gaps as necessary).

For each family F_i, construct a corresponding profile HMM ($\lambda(F_i)$).

Objective is to assign a newly sequenced protein to one of the k families.

Then the likelihood (\$P[O $|\lambda(F_i)]$) of the gap-aligned new protein is computed for each of the k profile HMMs. The new protein is then assigned to the family for which the likelihood is maximum. (The "scoring" problem of HMMs from last time.)