CSE 527 Lecture Notes
Lecture 19, Dec. 3, 2003
Tin Louie – tinlouie@u

First topic: HMMs in action

“Profile hidden Markov models can be used to do sensitive database searching using statistical descriptions of a sequence family’s consensus” (quote from http://pfam.wustl.edu/hmmsearch.shtml)

No structural similarity is taken into account ... this is NOT a method which includes scoring terms for nearby aa (amino acids) in a 3D structure.

Note that multiple alignment alone doesn’t cut it: (the following is a quote from http://hmmer.wustl.edu/) ”You have carefully constructed a multiple sequence alignment [on a protein sequence family]. Your family, like most protein families, has a number of strongly (but not absolutely) conserved key residues, separated by characteristic spacing. You wonder if there are more members of your family in the sequence databases, but the family is so evolutionarily diverse, a BLAST search with any individual sequence doesn’t even find the rest of the sequences you already know about. You’re sure there are some distantly related sequences in the noise. You spend many pleasant evenings scanning weak BLAST alignments by eye to find ones with the right key residues are in the right places, but you wish there was a computer program that did this a little better.”

Note that weight matrices don’t handle variable-length gaps.

The model: (see the lecture slides)
match states: emission
insert states: emission
...match portions of query sequences that do not match conserved regions
...follows a geometric distribution
delete states: no emissions

Transition & emission probabilities are based on counts from the training set.

Deciding between a match or an insert state:
... simple approach: if the multiple alignment shows that more than half of the sequences have a gap in this position, then decide this is an insert state
...better approach: maximum a posteriori assignment with forward-like algorithm

Searching: use forward algorithm or Viterbi algorithm
Alignment: Viterbi algorithm to calculate the most probable path
Scoring: log-likelihood or log-odds (log-odds: log of the ratio of the likelihoods of this model vs background/”random” model)

either scoring method can be converted to Z scores (the following is paraphrasing from ”Hidden Markov Models in Computational Biology”, Anders Krogh, Michael Brown, et al., J Mol Biol (1994))

” — quantify the difference between log-likelihood scores for proteins containing the domain & scores for proteins not containing the domain — using a local windowing technique, we first calculate a smooth average curve for the roughly linear band of the log-likelihood score vs length plot. The standard deviation around this average curve is also calculated. Using this, we calculate the difference between the scores of a sequence and the average score of typical sequences of that same length, measured in standard deviations. This number is called the Z-score for the sequence. We then choose a Z-score cut-off, either a priori or by looking at the histogram of Z scores, and use it to decide if a given sequence fits the model or not.”

Refinements:
for small training sets, use pseudocounts to avoid a count of zero and zero emission probability
... a constant (called A in the formula on the lecture slides) and background rate
... weighted mixtures of pseudocounts, based on probabilities of being in certain regions

downweight highly similar sequences due to sampling bias (the training set really should contain ”independent” sequences)

Second topic: gene prediction

ORF (open reading frame): no termination codon ... long ORFs are good indications of gene

Codon bias is another indication:
... within a given species, synonym usage is biased (not all possible triplets are used evenly)
... Leu : Ala : Trp in uniform random sequence should appear in 6:4:1 ratios, but in reality these aa are used in ratios closer to 7:6.5:1

3 graphs from ”Codon preference and its use in identifying protein coding regions in long DNA sequences”, Nucleic Acids Res. January 1982
... 5th or 6th order Markov models, all 3 reading frames
... predicting genes between stop codons
... overlapping genes on opposite strands are hard to find

Eukaryote genes are more complex:
... existence of introns, which also have ORFs (but are not part of the coding portion of the gene) and are highly variable in length
... splice donor / acceptor sites
... alternative splicing
... poly-A tail location variability
... 5' and 3' UTR (untranslated regions)