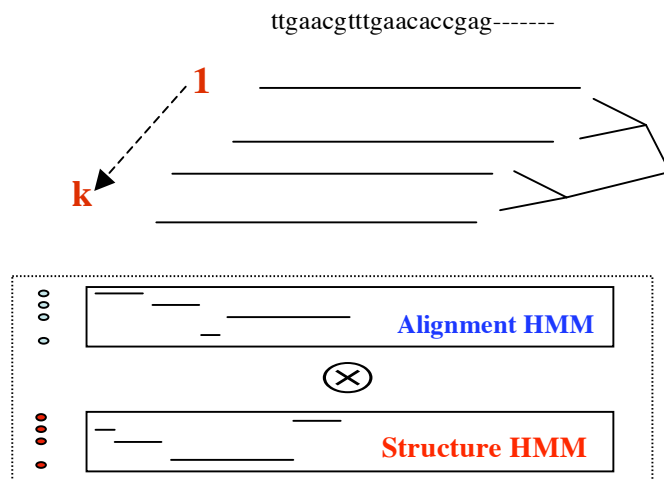


Simultaneous Alignment and Annotation

1.10.07

Annotation is clearly a central problem in comparative genomics: You use observed molecular evolution to make statements about something you can't observe, such as a folding RNA molecule or the protein gene structure in a genome segment. Alignment is more a nuisance problem that must be solved before interesting data analysis can be performed. The last decade has seen the rise of statistical alignment based on stochastic models of insertions, deletions and substitutions. Optimally, alignment and annotation should be solved together and this has been done in several publications for special cases. However, this is difficult to do properly for statistical alignment, but decent ad hoc methods has been proposed (Hobolt and Jensen, 2005; Satija, Pachter and Hein, 2008). Both these methods have a common underlying structure that is described in the illustration below. The statistical alignment can be solved via an HMM as for instance described in Lunter et al. (2005). If there is no alignment problem, then annotation can also be solved by an HMM. However, annotation based on a modelling combining stochastic models of insertions, deletions and substitutions with dependency on an annotation has not so far been shown to be solvable by an HMM. The problem stems from the fact that hidden structures such as exons would have an equilibrium distribution and time dynamics, that probably isn't describable by an HMM. If an observable sequence with its hidden states experience insertions and deletions, then the distribution of the hidden states will most likely stop being Markovian. The ad hoc methods above just combine the 2 HMMs anyway and it probably good for most purposes. This approach could be applied generally and developing software that combined structure HMMs with statistical alignment HMMs could have great value. The number of states in such combined HMMs could for many sequences and complex annotations be very large. However, just having the ability to triple (possibly quadruply) align could if combined with the idea in the project "Intermediates between Spanning and Steiner Trees" allow the annotation and alignment of an arbitrary number of sequences, which clearly would be an accomplishment and useful in face of the large number of coming genomes.

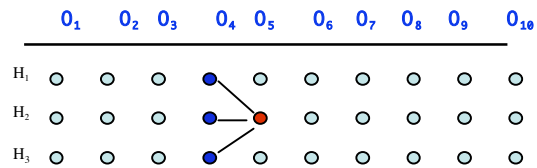


This approach to annotation-alignment consists of a combined HMM: The Structure HMM describes for instance genes structure, while alignment HMM describes the probability of different alignments under for instance the TKF91 model. Combined HMM will then have states corresponding to any possible alignment state and annotation. Above the HMMs is shown a phylogeny (tree) relating 4 sequences. The molecular evolution of a given position in the 4 sequences can be dependent on the chosen hidden state.

If we let HS and HA be the structure and alignment HMMs respectively. To illustrate using HS, let p_{ij} be probability of jumping between hidden neighbouring states i and j . π_k is the equilibrium probability of state k . $P(e_k=O_j)$ is the probability of emitting O_j if the hidden state is e_k . The probability of observing the first k columns given that the hidden state at column k

is j obeys the recursion below that is illustrated on a case with 10 observations and 3 hidden states behind each column.

$$P_{O_k=i}^{HS_k=j} = P(O_k = i | HS_k = j) \sum_{HS_{k-1}=r} P_{O_{k-1}=r}^{HS_{k-1}=r} p_{r,i}$$



The recursion to the left, states that the probability of the observations from column 1 to column k given the k 'th hidden state is j , can be partitioned into possible transitions between the $k-1$ 'st and the k 'th hidden state and observations from 1 to $k-1$, conditioned on the hidden states at $k-1$. This has to be multiplied with the probability of what was observed at k conditioned on the hidden state there. In the exemplification to the right k is 5 and hidden state there is 2. This recursion allows summation over all combinations of hidden states. Very similar algorithms can determine combinations of most probable hidden states.

Similar recursions can be written for alignment HMMs. To combine two HMMs, let $\{S_i\}$ be the structure states and $\{A_k\}$ the alignment states. There are a series of difficult decisions to be made in combining HMMs as can be seen in the project description "Artifacts from Combining Hidden Markov Models" at <http://mathgen.stats.ox.ac.uk/bioinformatics/projects/>.

There are different ways to combine, but the most straightforward would be to let $\{AS_m\}$ be $\{A_k, S_i\}$, and then $p(A_k, S_i \rightarrow A_{k'}, S_{i'}) = p(A_k \rightarrow A_{k'}) * p(S_i \rightarrow S_{i'})$ [normalized] and let $e_O(A_k, S_i) = e_O(A_k) e_O(S_i)$ [normalized]. Restrictions can be placed on AS, $p(A_k, S_i \rightarrow A_{k'}, S_{i'})$ and $e_O(A_k, S_i)$ based on modelling considerations as this description allow all combinations.

Could one introduce second order effects, ie not combine the alignment HMM and the structure HMM in so direct fashion, but investigate the effect of evolution and then propose an approximating HMM possibly of higher order?

Project Plan

- i. Analyze TKF91 (Lunter et al., 2004) and gene annotation (Pedersen et al., 2003) in a combined model for 2 sequences.
- ii. The same as i. for 3 and 4 sequences.
- iii. Simulate the evolution of a large number of pairs of sequences and investigate the fit of naïve combination of structure and alignment HMMs. Would a higher order HMM fit better?

And if time permits:

- iv. Analyse the situation of n sequences in general and consider the idea described in project "Partial and Stochastic Summation of Hidden States" to accelerate computations.

A natural extension of this would be to combine stochastic context grammars with statistical alignment.

References

- Durbin, Eddy, Krogh and Mitchison (1998) "Biological Sequence Analysis" Cambridge University Press
- Hobolth, A. and Jensen, J.L. (2005). Applications of hidden Markov models for characterization of homologous DNA sequences with a common gene. *Journal of Computational Biology*, **12**, 186-203.
- Lunter, Miklos, Drummond, & Hein (2005) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen.)
- Pedersen, J.S. and J.J. Hein (2003) "Gene finding with a hidden Markov model of genome structure and evolution" *Bioinformatics* 19.2.219-227.
- Rahul, Paether and Hein (2008) "Combining Statistical Alignment and Footprinting" (in preparation)