Poster H-11

Hidden Markov Models for Gene Finding in Viral Genomes

Authors:
Saskia de Groot (*Department of Statistics, University of Oxford*)
Stephen McCauley (*Department of Statistics, University of Oxford*)
Jotun Hein (*Department of Statistics, University of Oxford*)

Short Abstract: Viruses tend to code in overlapping reading frames. The constrictions forced upon multiple coding regions will result in atypical both codon bias and sequence evolution. Exploiting these latter constraints we present three different gene-finding HMMs – single, pairwise and EHMM – to annotate viral genomes coding in unidirectional overlapping reading frames.

Long Abstract:

Introduction: Detecting genes in viral genomes is a complex task. Due to the biological necessity of them being constrained in length, viruses tend to code in overlapping reading frames. Since one amino acid is encoded by three nucleic acids, up to three genes may be coded for simultaneously in one direction. Conventional HMM gene finding algorithms, when applied to viral genomes, may find it difficult if not impossible to identify multiple coding regions, since in general their topologies do not allow for the presence of overlapping or nested genes. The constraints forced upon the multiple coding regions by single nucleotides being involved in the coding of up to three genes simultaneously will result in an atypical both codon bias and sequence evolution. Exploiting these latter effects we present two different gene-finding programs to annotate viral genomes coding in unidirectional overlapping reading frames.

Firstly we introduce a method which makes use of distinct nucleotide signatures within genes and non-genes, in particular the predisposition towards the usage of highly degenerate codons in overlapping coding regions. Heuristic methods, in particular by Pavesi et al., have been used before to propose regions of certain sequence composition, e.g. high degeneracy or certain nucleotide motifs, as multiple coding. This has however as yet not been translated into a probabilistic HMM framework. We therefore first develop a self training single sequence annotation method comprising up to third order Markov Models which allows for overlapping reading frames. The emission of nucleotides is dependent on up to the two ones already emitted beforehand, thus being able to account for the observed codon bias in overlapping regions. All emission parameters are free so that additional features such as low GC content in coding regions may also be accounted for. Next we extend this single sequence annotation HMM (SSA) to an evolutionary EHMM as introduced by Siepel & Haussler. We use an alignment of homologous sequences related by a phylogenetic tree topology of unknown branch lengths and the Hein & Stoevlbaek evolutionary model. Iteratively estimating our model parameters, we update our single sequence and evolutionary model parameters and the phylogenetic distances simultaneously.

Secondly we introduce a pairwise HMM which uses solely the evolutionary conservation between two viral ssRNA sequences descended from a common ancestor to simultaneously annotate both. Evolutionary information, notably by Firth et al. has been used in prior

comparative methodologies to discover overlapping reading frames using simple likelihood ratio tests to compare the single coding to multiple coding hypotheses. However the evolutionary behaviour of overlapping reading frames has up till now not been integrated into a pair HMM methodology. For an alignment of two homologous viral genomes we therefore develop a pair HMM which incorporates the possibility of coding in multiple reading frames. Usually pair HMM methodologies are constrained to assuming conserved gene structure. Our model allows for a shift in Start and Stop codons having occurred and thus allows for gene structure to have evolved over time. The same evolutionary model as above is used and parameters are estimated using expectation maximization, thus providing an estimate of evolutionary distance as well as selection factors on certain genes.
We additionally demonstrate how to include biological knowledge - both true and false - into our models in form of a weighted prior on certain states. We demonstrate how this affects our predictions and discuss different possibilities of sensible priors and their merits.
Results: We successfully annotate a set of HIV2 strands up to a high level of sensitivity and specificity using the SSA method achieving average sensitivity of around 90% and specificity of 99%. The pairwise HMM method provides slightly lower average values being dependent on sequences being sufficiently divergent The jump from the use of the single sequence annotator to the EHMM greatly improves annotation power to an average sensitivity of around 96% and specificity of 98%, since it combines sequential and evolutionary information. However there are not always sufficient genomic sequences of adequate evolutionary distance within the same family available. We also used the pair HMM method to perform a comparative simultaneous annotation on an alignment of two strands of HIV1 and HIV2 resulting in prediction accuracy of 77% sensitivity and 98% specificity - since their gene structures have evolved quite substantially over time one could not use an HMM which insists on conserved gene structure. Additionally the SSA method was used to annotate every linear ssRNA virus listed in GenBank. As a result a number of novel regions, notably in Avian Leucosis, Primate T-lymphotronic 3 and Murine type C retro viruses, were annotated as coding.
Conclusion: Up to date only heuristic methodologies have been available for dealing with sequences coding in overlapping reading frames, a feature especially common in viral genomes. A self training HMM methodology which does allow for multiple coding regions is therefore invaluable as a tool for future work. We have developed several HMM frameworks suitable to different problems and applied these to a large data set accurately predicting genome structures as complex as HIV and additionally highlighting putative coding regions as of yet unannotated in GenBank. We conclude that this is a vital step towards understanding the evolution of viral sequences and their genome structure.

S. McCauley, J. Hein (2006) ``Using HMMs and observed evolution to annotate viral genomes'' {Bioinformatics} online on April 13, 2006

A.E. Firth, C.M. Brown (2005) ``Detecting overlapping coding sequences with pairwise alignments'' {Bioinformatics}, 21(3), 282-292

A.E. Firth, C.M. Brown (2006) ``Detecting overlapping coding sequences in virus genomes'' {BMC Bioinformatics}

A. Pavesi, B. De Iaco, M.I. Granero, A. Poratei (1997) ``On the informational content of

overlapping genes in prokaryotic and eukaryotic viruses'' {Journal of Molecular Evolution}, 44(6), 625-631

A. Pavesi(2000) ``Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus'' {Journal of Molecular Evolution}, 50(3), 284-295

A. Siepel, D. Haussler (2004) ``Combining phylogenetic and hidden Markov models in biosequence analysis'' {Journal of Computational Biology}, 11(2-3), 413-428