Sequence analysis

Using hidden Markov models and observed evolution to annotate viral genomes

Stephen McCauley* and Jotun Hein
Department of Statistics, Oxford University, Oxford, UK

Received on August 19, 2006; revised and accepted on March 9, 2006
Advance Access publication April 13, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: ssRNA (single stranded) viral genomes are generally constrained in length and utilize overlapping reading frames to maximally exploit the coding potential within the genome length restrictions. This overlapping coding phenomenon leads to complex evolutionary constraints operating on the genome. In regions which code for more than one protein, silent mutations in one reading frame generally have a protein coding effect in another. To maximize coding flexibility in all reading frames, overlapping regions are often compositionally biased towards amino acids which are 6-fold degenerate with respect to the 64 codon alphabet. Previous methodologies have used this fact in an ad hoc manner to look for overlapping genes by motif matching. In this paper differentiated nucleotide compositional patterns in overlapping regions are incorporated into a probabilistic hidden Markov model (HMM) framework which is used to annotate ssRNA viral genomes. This work focuses on single sequence annotation and applies an HMM framework to ssRNA viral annotation. A description of how the HMM is parameterized, whilst annotating within a missing data framework is given. A Phylogenetic HMM (Phylo-HMM) extension, as applied to 14 aligned HIV2 sequences is also presented. This evolutionary extension serves as an illustration of the potential of the Phylo-HMM framework for ssRNA viral genomic annotation.

Results: The single sequence annotation procedure (SSA) is applied to 14 different strains of the HIV2 virus. Further results on alternative ssRNA viral genomes are presented to illustrate more generally the performance of the method. The results of the SSA method are encouraging however there is still room for improvement, and since there is overwhelming evidence to indicate that comparative methods can improve coding sequence (CDS) annotation, the SSA method is extended to a Phylo-HMM to incorporate evolutionary information. The Phylo-HMM extension is applied to the same set of 14 HIV2 sequences which are pre-aligned. The performance improvement that results from including the evolutionary information in the analysis is illustrated.

Availability: We implement the SSA method in the MATLAB programming language and provide the source code at http://www.stats.ox.ac.uk/Qmccauley. Additional supplementary material referred to in the text is available on the same webpage.

Contact: mcccauley@stats.ox.ac.uk

Supplementary Information: Supplementary data are available at http://www.stats.ox.ac.uk/Qmccauley

*To whom correspondence should be addressed.

1 INTRODUCTION

This paper presents a hidden Markov model (HMM) topology designed to allow for overlapping CDS regions. This HMM topology is incorporated into an SSA program which is used successfully to annotate ssRNA viral genomes. Overlapping CDS regions are of interest because there are multiple evolutionary coding constraints operating on them. These multiple constraints complicate any statistical inference procedure that might be applied to the genomic analysis of ssRNA viral genomes.

This work focuses on annotating ssRNA viral genomes because they are known to harbour overlapping CDS regions in relative abundance. The issues of modelling the evolutionary constraints on these regions, and making general inference as to their abundance, is confounded by the fact they are difficult to identify. This is often because they are small and nested within longer CDS regions in alternative open reading frames (ORFs). This is illustrated by the fact that new overlapping CDS regions within viral genomes are still being identified (Brocchieri et al., 2005). Whilst existing CDS-finding methodologies such as GeneMarkS (Besemer et al., 1999, 2001) have been successful in identifying the majority of ssRNA viral CDS regions, it remains probable that the identification of all overlapping CDS regions within viral genomes is incomplete. It is of obvious biological importance to identify any additional CDS within viral genomes that have been missed, and so any methodologies likely to further such identification are well motivated. When an evolutionary model is included to improve the viral annotation, the estimates of the parameters of the evolutionary model are themselves interesting and worthy of consideration.

There are very many documented patterns of overlaps which fall into the three main classes of unidirectional (CDS overlap in same reading direction), convergent (CDS 3' ends overlap in opposing reading directions) and divergent (CDS 5' starts overlap in opposing reading directions). Some examples of these include Larrey et al. (1996), Shmulevitz et al. (2002), Walewski et al. (2001) and Zajanckauskaite et al. (1997). Study has been made into these overlapping regions and the nature of the evolutionary pressures that they are subject to (Guyader and Ducray, 2002; Hughes et al., 2001; Mizokami et al., 1997; Rogozin et al., 2002). The success of the single sequence annotation (SSA) methodology presented in this paper is based on an HMM topology which allows for any given nucleotide to code in multiple reading frames as well as allowing for differentiated nucleotide compositional patterns in overlapping CDS with respect to non-overlapping CDS and non-CDS regions. These differentiated nucleotide
compositional patterns are a result of the particular evolutionary pressures that overlapping CDS are subjected to, some of which are described in the papers cited above. Pavesi et al. (1997) and Pavesi (2000) exploited information on common overlapping CDS motifs to search for as yet unknown overlapping CDS with success. More theoretical work modelling overlapping CDS includes Kozlov (2000a, b) and Krakauer (2000).

There has been some research in the identification and evolution of overlapping CDS but it remains an understudied area. Recently Firth and Brown (2005) proposed a methodology for analysing overlapping CDS with pairwise alignments where they model differentiated mutation rates between aligned overlapping CDS and non-overlapping CDS. Hein and Stövlbæk (1995) also investigated evolutionary modelling of overlapping CDS and applied their methods to a pairwise alignment of HIV sequences. Further work on estimating evolutionary parameters between pairwise alignments was undertaken by Pedersen and Jensen (2001). These papers clearly illustrate that evolutionary methods are useful for analysing overlapping CDS. Although Firth and Brown (2005) refer to HMM methods as being complex to model and difficult to parameterize, this work has employed an HMM framework and has achieved significant success at annotating overlapping CDS in single sequences.

As a first step towards an evolutionary extension of this model, a Phylogenetic-hidden Markov model (Phylo-HMM) framework has been developed and applied to an alignment of 14 HIV2 sequences using a parameterized evolutionary model. The potential for improved annotation based on the additional evolutionary signal within the Phylo-HMM framework is illustrated. Because these methodologies are applied at the nucleotide level, there is the potential to extend them to account for additional evolutionary constraints imposed upon the ssRNA nucleotides (such as secondary structure conservation). This has already been considered (albeit from a different perspective) by Pedersen et al. (2004a) who examined differentiated mutation rates in coding and non-coding regions depending upon the secondary structure constraints imposed upon the nucleotide (Pedersen et al., 2004b). Their methods only consider non-overlapping coding regions (whose CDS annotation is known), and a combination of approaches may serve to give a more generalized and accurate model of the evolutionary mutational processes operating on ssRNA viral genomes thereby improving both the CDS annotation and secondary structure predictions.

2 METHODOLOGY

The SSA method takes as input a genomic sequence and annotates the genome into regions of CDS and non-CDS within the HMM framework discussed below. The method parameterizes this HMM concurrently with the annotation in an iterative Expectation Maximization procedure. For a given set of parameters, the method maximizes with respect to an annotation, the overall likelihood of observing the genome. This is accomplished using Viterbi’s procedure (Durbin et al., 1998). The essence of the HMM framework is that the observed nucleotides are distributed according to a probability distribution that depends on the coding characteristics of the various reading frames. In order to describe this particular HMM, it is necessary to describe the topology of the HMM by defining the various states and emission states and how they are related. A detailed discussion of the transition and emission parameters as well as a description of how the SSA procedure is implemented follows.

2.1 Defining the HMM topology

The HMM topology is designed to accommodate unidirectional overlaps which are the most common form of overlapping CDS regions. This methodology allows for any given nucleotide to code in up to a maximum of three overlapping frames reading in one direction. By making certain assumptions regarding the emission distributions of the nucleotides we can accommodate an overlapping CDS model in eight states. The HMM topology is defined by describing the states in the HMM, the possible transitions between these states and the emission distributions from each of these states.

2.1.1 HMM states

The States are illustrated in Figure 1.

(1) State NC: Describes a nucleotide which is not coding in the genome.
(2) State S 1: Describes a nucleotide coding at the first codon locus in a codon within a region of the genome encoding only one CDS.
(3) State S 2: Describes a nucleotide coding at the second codon locus in a codon within a region of the genome encoding only one CDS.
(4) State S 3: Describes a nucleotide coding at the third codon locus in a codon within a region of the genome encoding only one CDS.
(5) State D 1,2: Describes a nucleotide coding at the first codon locus in a codon in one reading frame and at the second in another within a region of the genome coding in two distinct reading frames.
(6) State D 2,3: Describes a nucleotide coding at the second codon locus in a codon in one reading frame and at the third in another within a region of the genome coding in two distinct reading frames.
(7) State D 3,1: Describes a nucleotide coding at the third codon locus in a codon in one reading frame and at the first in another within a region of the genome coding in two distinct reading frames.
(8) State T 1,2,3: Describes those rare nucleotides coding at the first, second and third codon loci in three distinct reading frames.

2.1.2 HMM state transitions

For the state transitions again please refer Figure 1. Those transitions leaving each of the eight states are described, two for each state, giving a total of 16 labelled as T1–T16. The single parameter τ represents the probability of entering a CDS region conditional on the previous three emissions coding AUG (a potential START codon).

(1) Transitions from State NC will generally be the self-transitioning cycle (T1). In the case where the previous three emissions are AUG there is a probability τ that the HMM will also make a transition to State S 1 (T2).
(2) Transitions from State S 1 will generally be to S 2 (T3) but if the previous three nucleotides were AUG then the HMM goes to state D 1,2 with
Table 1. Summarising the conditional emission distributions

<table>
<thead>
<tr>
<th>Conditional emission distributions</th>
<th>Nucleotides upon which to Condition</th>
<th>Number of conditional permutations</th>
<th>Maximum number of parameters</th>
<th>SSA number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CED_{NC}</td>
<td>none</td>
<td>0</td>
<td>1+3</td>
<td>1+3</td>
</tr>
<tr>
<td>CED_{1}</td>
<td>none</td>
<td>0</td>
<td>1+3</td>
<td>1+3</td>
</tr>
<tr>
<td>CED_{2}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
<tr>
<td>CED_{3}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
<tr>
<td>CED_{D_{1},1}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
<tr>
<td>CED_{D_{1},2}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
<tr>
<td>CED_{D_{1},3}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
<tr>
<td>CED_{D_{1},4}</td>
<td>(e_{i-1}, e_{i-2})</td>
<td>16</td>
<td>4+3</td>
<td>4+3</td>
</tr>
</tbody>
</table>

CED_{NC} pertains only to State NC; CED_{1}, CED_{2} and CED_{3} pertain to the single CDS states; CED_{D_{1},1}, CED_{D_{1},2} and CED_{D_{1},3} similarly describe three periodic paired codon loci in the reading frames of the double CDS states; CED_{T_{1,2,3}} describes the triplet CDS State \(T_{1,2,3}\). \(e_{i-1}\) represents the previously emitted nucleotide in the sequence. Note the difference between the maximum number of parameters in the most highly parameterised model possible using this modelling criteria, and the actual number of parameters used in the SSA procedure where the multiple coding emission parameters are derived from the singly coding parameters.

(3) Transitions from State \(S_{1}\) will generally be to \(S_{5}\) (T5) but if the previous three nucleotides were AUG then the HMM transitions to state \(D_{3,1}\) with probability \(1\) (T6).

(4) Transitions from State \(S_{2}\) will generally be to \(S_{1}\) (T7) but if the previous three nucleotides represent a STOP codon then the HMM transitions to the NC state with unitary probability \(1\). Even if the previous three nucleotides encode an AUG there is no transition to a double CDS state since there already exists a CDS in this reading frame.

(5) Transitions from State \(D_{1,2}\) will generally be to \(D_{2,3}\) (T9) but if the previous three nucleotides were AUG then the HMM transitions to the triplet coding state \(T_{1,2,3}\) with probability \(1\) (T10). Since neither of the CDS regions has emitted their third codon locus nucleotide, there is no possibility of terminating either of the CDS regions.

(6) Transitions from State \(D_{2,3}\) will generally be to the \(D_{3,1}\) state (T11), but if the previous three nucleotides represent a STOP codon then the HMM transitions to \(S_{3}\) (T12).

(7) Transitions from State \(D_{3,1}\) will generally be to the \(D_{1,2}\) state (T13), but if the previous three nucleotides represent a STOP codon then the HMM transitions to \(S_{2}\) (T14).

(8) Transitions from State \(T_{1,2,3}\) will generally be the self-transitioning cycle (T15) but if the three previous nucleotides represent a STOP codon then the HMM transitions to \(D_{2,3}\) (T16).

2.1.3 HMM conditional emission distributions

There are eight conditional emission distributions (CEDs) to define within the SSA procedure each corresponding to one to one with the eight states. Table 1 illustrates the nature of these emission distributions.

(1) CED_{NC}: Describes the emissions in non-coding regions which are modelled by sampling from a multinomial distribution across the four-letter nucleotide alphabet.

(2) CED_{1}: Describes the first three periodic coding locus in a CDS codon. In this methodology the level of conditioning is only according to the present codon so in fact this CED_{1} can be described as sampling from a multinomial distribution across the four-letter nucleotide alphabet.

(3) CED_{2}: Describes the second three periodic coding locus in a CDS codon. This CED_{2} can be represented as conditional sampling from one of four multinomial distributions across the four-letter nucleotide alphabet.

(4) CED_{3}: Describes the third three periodic coding locus in a CDS codon. This CED_{3} can be represented as conditional sampling from one of 16 multinomial distributions across the four-letter nucleotide alphabet.

(5) CED_{D_{1},1}: Describes the first three periodic coding locus in one CDS codon and the second coding locus in another. CED_{D_{1},1} can be described by conditional sampling from one of four multinomial distributions across the four-letter nucleotide alphabet.

(6) CED_{D_{1},2}: Describes the first three periodic coding locus in one CDS codon and the third coding locus in another. CED_{D_{1},2} can be represented as conditional sampling from one of 16 multinomial distributions across the four-letter nucleotide alphabet.

(7) CED_{D_{1},3}: Describes the second coding locus in one CDS codon and the third coding locus in another CDS. CED_{D_{1},3} can also be represented as conditional sampling from one of 16 multinomial distributions across the four-letter nucleotide alphabet.

(8) CED_{T_{1,2,3}}: Represents the first coding locus in one CDS and the second coding locus in another CDS and the third coding locus in the final CDS. CED_{T_{1,2,3}} can be represented as conditional sampling from one of 16 multinomial distributions across the four-letter nucleotide alphabet.

2.2 Expectation maximisation joint parameterisation and annotation procedure

Viral genomes differ substantially in their overall nucleotide composition, therefore it is unrealistic to expect one set of parameters to best describe all viral genomes. Thus it is required to parameterize the HMM according to the specific viral genome at hand.

If \(P\) is the multidimensional parameter space, and \(A\) is the annotation space, then for every set of parameters \(P\) there is a maximally likely annotation \(\hat{A}\) as calculated using Viterbi. The stepwise Baum–Welch iterative procedure’s convergence to the global maximum \(P_{\text{max}}\) (with an associated set of parameters \(P_{\text{max}}\)) is dependent on the global properties of the multidimensional likelihood function and its smoothness as a function of \(P\).

Related to these points is the distance of \(P_{\text{max}}\) from \(P_{\text{init}}\). If the likelihood function is not a smooth function of \(P\) then there may be local maxima where the EM procedure may get stuck and the closer \(P_{\text{init}}\) to \(P_{\text{max}}\), the less likely this is to happen.

The first step in the SSA procedure is feeding the SSA algorithm some seed HMM parameters (which are derived from a seed annotation). At first every ORF of nucleotide length greater than some threshold (typically 200 nt) is labelled as coding and this seed annotation is used to calculate the Maximum Likelihood estimates of the HMM parameters accordingly (the convergence of the Baum–Welch training procedure is quite robust to this threshold value). The most likely estimates of the parameters are identified by summing over possible state annotations in a probabilistic fashion (using the old estimates of the parameters). The old parameters are updating in an iterative loop until the change in the probability of the data falls below a certain threshold. The most likely state annotation is then calculated using Viterbi’s algorithm with these maximally likely parameters estimates.

One difficulty parameterizing according to this method is that any initial seed annotation is unlikely to be correct, and will undoubtedly render more information regarding non-CDS and single CDS regions than overlapping CDS regions. It is a point of concern that the nucleotide composition of overlapping regions is likely to vary too much in the iterative procedure because they typically cover a smaller percentage of the genome and ssRNA
viral genomes in general tend to be short, therefore there is little data to train on. Various strategies were tested including utilizing Bayesian priors for the overlapping nucleotide compositions. The simplest approach, namely deriving the overlapping nucleotide compositional emission parameters as an arithmetic average of the relevant single coding emission parameters, performed on par with more complex prior models. Consequently this is the method implemented in the SSA procedure. Since the singly coding parameters are updated in successive Baum–Welch iterations, so too are the derived overlapping emission parameters. This allows for differentiated nucleotide compositions in overlapping regions, but does not necessarily skew the overlapping emission probability towards emitting a higher overall fraction of degenerate codons (which we may or may not suspect a priori). There are means of skewing the parameters in such a manner (such as weighting the third codon positions lower in the arithmetic averaging procedure), but this involves introducing another set of parameters (weights) and again in the short genomes examined this did not seem to affect the overall probability or annotation significantly.

In the scenario where a viral genome has several triple coding regions the SSA model will fit the same emission distribution to each of these regions (because they are assigned the same state in the HMM topology). However there might be particular local phenomenon which perturb the emission distributions of these regions so they are in fact quite different from a nucleotide compositional perspective. They will be modelled as having the same emission distribution in this SSA model, and to the extent that this is incorrect the accuracy of the model and therefore the overall annotation will suffer. The SSA method has limited data to train on and while there are known scenarios which it does not model, it nevertheless generates some encouraging results.

2.3 An illustration of a phylogenetic-HMM extension to the SSA methodology

The SSA method has been extended to incorporate comparative information via a Phylo-HMM. Phylo-HMMs incorporate the time dimension into the model by utilizing another set of Markov processes which describe how each genomic locus tends to evolve through time (Siepel and Haussler, 2004). Phylo-HMMs allow for the capture of both nucleotide sequence and nucleotide evolution information and these can be related back to the underlying hidden state of the HMM. This particular Phylo-HMM can be understood in terms of extending the single sequence HMM which emits a single nucleotide at any locus, to the case where this emitted nucleotide then generates the other nucleotides in the alignment according to a set of general time reversible (GTR) mutational models and a phylogenetic tree. For the mutational model we chose an extension of Hein and Støvlbæk’s implementation (see Table 4, last row, first numerical column) and the performance on one strain (GenBank Accession Number J04498) is illustrated in detail. The SSA predicted annotation for strain J04498 is illustrated in Figure 2B with the GenBank annotation above (Fig. 2A) and the extended Phylo-HMM prediction below (Fig. 2C). A cursory comparison indicates general agreement between the GenBank and the SSA annotations which is in line with the overall sensitivity of 93.1% (Table 4 second row first numeric column) and the performance of accuracy in the three reading frames would be desirable. The measurement of sensitivity and specificity quoted are based on an effective coding genome size where nucleotides which code in two reading frames count twice in the sensitivity measurement and so on. The denominator of the specificity statistic is the sum over the entire genome of the number of CDS reading frames each nucleotide codes in. The numerator is the sum of the correctly predicted CDS reading frames each nucleotide codes in. The denominator of the specificity statistic is the sum over the genome of the number of correctly predicted non-coding reading frames for each nucleotide. It should be noted that in the case where the model correctly predicts a new CDS region that is not in the GenBank database, this will show up as a decrease in the specificity measurement.

3 RESULTS

3.1 SSA as applied to HIV2 genomes

The SSA method was applied to 14 strains of HIV2. The SSA method had an average sensitivity of 90% overall on these genomes (see Table 4, last row, first numerical column) and the performance of accuracy in the three reading frames would be desirable. The measurement of sensitivity and specificity quoted are based on an effective coding genome size where nucleotides which code in two reading frames count twice in the sensitivity measurement and so on. The denominator of the specificity statistic is the sum over the entire genome of the number of CDS reading frames each nucleotide codes in. The numerator is the sum of the correctly predicted CDS reading frames each nucleotide codes in. The denominator of the specificity statistic is the sum over the genome of the number of correctly predicted non-coding reading frames for each nucleotide. It should be noted that in the case where the model correctly predicts a new CDS region that is not in the GenBank database, this will show up as a decrease in the specificity measurement.
likely HMM path, it is possible to analyse these regions and consider the mean probability that they are indeed correctly annotated. This is accomplished using Posterior Decoding (Durbin et al., 1998) at the nucleotide level and averaging over all nucleotides in that given region. These regions are represented as intervals on a line parallel to the x-axis in Figure 2B and the mean posterior decoding probabilities for these regions are indicated by arrows pointing to these regions. The two small incorrectly predicted CDS regions observed in gRF1 have mean posterior probabilities of 0.6 and 0.55 (which may give some indication that they are predictions not to be relied upon). Those nucleotide loci of the missed CDS regions have very high posterior probabilities even though the annotation is incorrect which indicates that the modelling procedure is not sufficiently sensitive to detect these incorrect annotations.

3.2 SSA as applied to a selection of ssRNA genomes

The SSA method has been applied to the annotation of other ssRNA viral genomes and the following subset of nine illustrate the successes and limitations of the methodology. The subset of nine genomes are as follows:

(1) Marburg: Marburg of the family Filoviridae. Sequenced by Feldmann et al. (1992) and Bukreyev et al. (1995) this 19 112 nt genome does not contain overlapping CDS (although it does contain two overlapping gene products as delineated by conserved transcriptional signals, beginning with a start site at the 3’ genome end and a terminating transcriptional polyadenylation site) and so represents the least complex genome to annotate from the perspective of CDS overlaps.

(2) Ebola: Reston Ebola of the family Filoviridae. Sequenced and analysed by Sanchez et al. (1993, 1996), Volchkov et al. (1999) and Grosseth et al. (2002) this 18 891 nt genome contains a region which encodes for two proteins. The smaller sGP protein starts and terminates in the same gRF, but the longer GP protein has the same start point but ends in a different gRF (this is accomplished via a process known as transcriptional editing).
(3) Hep G: Hepatitis G of the family Flaviviridae. Sequenced by Linnen et al. (1996) this 9392 nt genome contains only one CDS region which encodes a precursor polyprotein and from the perspective of this analysis represents a simple genome.

(4) EIAV: Equine infectious anemia virus is a lentivirus of the family Retroviridae. The genome we considered was published by Petropoulos (1997) and is 8359 nt long with two areas of overlapping CDS.

(5) SHIV: Simian-Human immunodeficiency virus is a lentivirus of the family Retroviridae. The genome we consider was published by Reimann et al. (1996) and is 10,000 nt long with 6 areas of CDS overlap and two CDS regions (tat and rev) which are further complicated by the presence of introns. This represents a difficult genome to annotate according to our methodology.

(6) Rous S: Rous Sarcoma is an Alpharetrovirus of the family Retroviridae. The genome we considered was again published in Petropoulos (1997) and the genome is 9392 nt long with 3 areas of overlapping CDS. There are several introns interspersing these CDS regions and this genome would be considered difficult to annotate.

(7) HIV2: Another lentivirus of the family Retroviridae. The genome we consider was published by Kirchhoff et al. (1990) and is similar in structure to Simian-Human immunodeficiency virus but with longer overlaps in the CDS.

(8) SHFV: Simian hemorrhagic fever virus is an arterivirus of the family Arteriviridae and the 15,717 nt genome we consider was sequenced and analysed by Smith et al. (1997), Godeny et al. (1995) and Zeng et al. (1995). The genome contains many CDS regions, introns and many overlaps. Many of the overlaps are short and the large number of them makes this a difficult genome to annotate.

(9) HTLV: Human T-lymphotropic virus is a deltaretrovirus of the family Retroviridae. The 8507 nt genome we consider is again published in Petropoulos (1997). This is quite a complex genome to annotate with multiple examples of ribosomal slippage and introns.

In this section the annotations of the SSA method are compared with those returned by the GeneMarkS annotation procedure (Besemer et al. 1999a). Comparisons of overall sensitivity and specificity measurements are illustrated, as are those measurements applied only to the overlapping regions (if any) in the applicable genome.

When submitting the genomes to GeneMarkS, since all genomes were <100,000 nt in length, the following program was automatically utilized: GeneMark.hmm PROKARYOTIC Version 2.2. The GeneMark.hmm program used to annotate the submitted genomes depends on heuristic methods to parameterize the HMM models. These heuristics relate the observed relationships between the positional frequencies and the global frequencies as well as relationships between the amino acid frequencies and the global GC percentage. The parameterized models are three periodicities for coding sequence of orders zero, one and two and a single zero order model for non-coding sequence [Besemer et al. (1999a)].

Table 2. Comparing both the overall and overlapping sensitivities of the SSA method with GeneMark.hmm

<table>
<thead>
<tr>
<th>Virus</th>
<th>SSA sensitivity</th>
<th>GMhmm sensitivity</th>
<th>SSA overlapping sensitivity</th>
<th>GMhmm overlapping sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marburg</td>
<td>1.0000</td>
<td>0.9816</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Ebola</td>
<td>0.9462</td>
<td>0.9334</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Hep G</td>
<td>0.9892</td>
<td>1.0000</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>EIAV</td>
<td>1.0000</td>
<td>0.9718</td>
<td>1.0000</td>
<td>0.0455</td>
</tr>
<tr>
<td>SHIV</td>
<td>0.9448</td>
<td>0.8739</td>
<td>0.8213</td>
<td>0.1156</td>
</tr>
<tr>
<td>Rous S.</td>
<td>0.9597</td>
<td>0.8545</td>
<td>0.9658</td>
<td>0.0000</td>
</tr>
<tr>
<td>HIV 2</td>
<td>0.8864</td>
<td>0.8245</td>
<td>0.4538</td>
<td>0.0009</td>
</tr>
<tr>
<td>SHFV</td>
<td>0.8912</td>
<td>0.7418</td>
<td>0.6063</td>
<td>0.0045</td>
</tr>
<tr>
<td>HTLV</td>
<td>0.8135</td>
<td>0.3492</td>
<td>0.1403</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The genomes were selected according to a rough measure of genomic complexity, from the very straightforward Marburg to the extremely complex HTLV.

Table 3. Comparing both the overall and overlapping specificities of the SSA method with Genemark.hmm

<table>
<thead>
<tr>
<th>Virus</th>
<th>SSA specificity</th>
<th>GMhmm specificity</th>
<th>SSA overlapping specificity</th>
<th>GMhmm overlapping specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marburg</td>
<td>1.0000</td>
<td>0.9979</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Ebola</td>
<td>0.9726</td>
<td>0.9765</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Hep G</td>
<td>1.0000</td>
<td>0.9927</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>EIAV</td>
<td>1.0000</td>
<td>0.9964</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>SHIV</td>
<td>1.0000</td>
<td>0.9868</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Rous S.</td>
<td>0.9857</td>
<td>0.9895</td>
<td>0.2194</td>
<td>n/a</td>
</tr>
<tr>
<td>HIV 2</td>
<td>0.9813</td>
<td>0.9875</td>
<td>0.7157</td>
<td>1.0000</td>
</tr>
<tr>
<td>SHFV</td>
<td>1.0000</td>
<td>0.9950</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HTLV</td>
<td>0.9948</td>
<td>0.9976</td>
<td>0.5737</td>
<td>n/a</td>
</tr>
</tbody>
</table>

The genomes were selected according to a rough measure of genomic complexity from the very straightforward Marburg to the extremely complex HTLV.

If we compare the overall sensitivities of the two methods in Table 2 (first two numeric columns) then we immediately observe that the overall sensitivities of the SSA method is generally superior to those predicted by GeneMark.hmm. In all but one genome the SSA model outperforms GeneMark.hmm, and in that case the results are within a percent or two so the performance is fairly similar. As with the majority of CDS finding methods the SSA method has a very high overall specificity as observed in Table 3 (first two numeric columns) and those high numbers compare favourably with Genemark.hmm.

If we compare the overlapping sensitivities in Table 2 (final two numeric columns) then the improvement that the SSA model brings to the analysis of ssRNA viral genomes is highlighted. (Note that these numbers are based on far fewer nucleotides so a large percentage difference can represent a fairly small number of correct/incorrect nucleotide annotations.) The first three genomes in the table do not have CDS encoded in different overlapping gRFs (therefore we would consider them to be non-complex from an...
Table 4. Overall and overlapping sensitivity comparisons for 3 of the 14 HIV2 sequences (only 3 are highlighted for ease of reading)

<table>
<thead>
<tr>
<th>HIV2 Strain accession number</th>
<th>SSA overall sensitivity</th>
<th>PHMM overall sensitivity</th>
<th>Difference in model sensitivities</th>
<th>SSA overlapping sensitivity</th>
<th>PHMM overlapping sensitivity</th>
<th>Difference in model sensitivities</th>
</tr>
</thead>
<tbody>
<tr>
<td>D00835</td>
<td>0.8851</td>
<td>0.9432</td>
<td>0.0581</td>
<td>0.6220</td>
<td>0.8699</td>
<td>0.2479</td>
</tr>
<tr>
<td>J04498</td>
<td>0.9308</td>
<td>0.9939</td>
<td>0.0631</td>
<td>0.6816</td>
<td>0.8670</td>
<td>0.1854</td>
</tr>
<tr>
<td>AY509260</td>
<td>0.8257</td>
<td>0.9101</td>
<td>0.0844</td>
<td>0.4618</td>
<td>0.7405</td>
<td>0.2787</td>
</tr>
<tr>
<td>MEANS OF 14</td>
<td>0.9010</td>
<td>0.9683</td>
<td>0.0673</td>
<td>0.6462</td>
<td>0.8644</td>
<td>0.2182</td>
</tr>
</tbody>
</table>

The last row is the overall mean results for the 14 sequences, and in general there is a 6.73% boost in overall sensitivity for reverting to the Phylo-HMM but there is a larger increase of 21.82% in the overlapping sensitivity for using the Phylo-HMM.

Table 5. Overall and overlapping specificity comparisons for 3 of the 14 HIV2 sequences (only 3 are highlighted for ease of reading)

<table>
<thead>
<tr>
<th>HIV2 Strain accession number</th>
<th>SSA overall specificity</th>
<th>PHMM overall specificity</th>
<th>Difference in model specificities</th>
<th>SSA overlapping specificity</th>
<th>PHMM overlapping specificity</th>
<th>Difference in model specificities</th>
</tr>
</thead>
<tbody>
<tr>
<td>D00835</td>
<td>0.9830</td>
<td>0.9733</td>
<td>−0.0097</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>J04498</td>
<td>0.9951</td>
<td>0.9928</td>
<td>−0.0023</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>AY509260</td>
<td>0.9928</td>
<td>0.9790</td>
<td>−0.0138</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>MEANS OF 14</td>
<td>0.9908</td>
<td>0.9811</td>
<td>−0.0092</td>
<td>1.0000</td>
<td>0.9918</td>
<td>−0.0082</td>
</tr>
</tbody>
</table>

The last row is the overall mean results for the 14 sequences, and in general there is a 0.92% loss in overall specificity for reverting to the Phylo-HMM and there is a comparable 0.82% loss in overlapping specificity for using the Phylo-HMM.

annotating perspective). There is an instance of transcriptional editing in the Ebola genome, where one smaller secreted non-structural glycoprotein sGP is encoded in one reading frame, and where a longer structural glycoprotein GP uses the same start sequence as sGP but its coding is extended beyond the end of the sGP CDS in a different reading frame by means of an inserted nucleotide at the transcriptional stage. The SSA model is not designed to pick up on such processes as transcriptional editing, but the model does flag both regions as coding, and indeed highlights these different products as partially overlapping CDS. From a perspective of overlapping CDS in different reading frames however, these first three genomes do not contain any, so they have no overlapping sensitivity statistics since there are no overlaps to identify. The SSA model does predict a small overlap however and so has a zero specificity (as does the GeneMark.hmm program) but this result is not as bad as it sounds, because by virtue of not predicting a small overlap, a large coding region that encodes the post-transcriptional editing site of the GP CDS would be missed entirely.

There is a sequential unidirectional overlap in Equine Infectious Anemia Virus of 200 or so nucleotides which only the SSA model annotates correctly, so from a sensitivity viewpoint the SSA model performs much better. Neither of the models predict any incorrectly labelled overlapping states but the specificity statistics are both unitary for EIAV. Simian-Human Immunodeficiency Virus is an even more complex virus with considerably more overlaps and a similar pattern is observed, in that the SSA model has a high sensitivity and low specificity, whereas GeneMark.hmm has low overlapping sensitivity. Rous Sarcoma is an interesting example where the SSA model again outperforms on a sensitivity basis, however the model has a low specificity which results from a prediction of 2 nested frames of 189 and 183 nt nested within the p60-SRC phosphoprotein CDS.

Of the remaining three genomes the SSA methodology clearly outperforms at predicting overlapping CDS. In the complex HIV 2 genome it predicts ~45% of the overlaps and in Simian Hemorrhagic Fever Virus it predicts ~60%. Human T-Lymphotropic Virus is an extremely complex genome with multiple overlaps and multiple examples of ribosomal slippage and introns which are exceptions to the behaviour captured in our model. Comparatively the SSA method performs well but there is still room for improvement in the more complex viral genomes.

3.3 Phylo-HMM applied to HIV2 genomes

The Phylo-HMM method is applied to each of the 14 HIV2 example sequences which were downloaded from GenBank and aligned with ClustalW [Thompson et al. (1994)]. Results are illustrated in Tables 4, 5 and Figure 2. A simple phylogeny was constructed using Felsenstein’s Phylip program [Felsenstein (1981)] which is illustrated in Figure 3. Each of the 14 sequences was annotated according to its own nucleotide composition and the observed evolution (the significance of which is interpreted using the mutational models and the phylogenetic tree).

The mutational models were parameterized using a manual iterative procedure which involved searching through the multidimensional parameter space for the Maximum Likelihood estimates of the multidimensional likelihood function. It would be ideal for these inputs to the Phylo-HMM to be calculated within the Phylo-HMM process, in which case it might be expected to improve
Annotating overlapping viral CDS

4 DISCUSSION AND FURTHER WORK

The results of the SSA methodology as applied to HIV2 ssRNA viral genomes are encouraging. Of the 14 strains of HIV2 examined in this work, the overall mean SSA sensitivity was $\sim 90\%$ which indicates that the method is performing well on average. The SSA mean specificity was $\sim 99\%$ which is further evidence that the method is performing well. The results of the SSA method on the nine ssRNA genomes indicate that this level of sensitivity and specificity are consistent with other results.

There is a limit to the accuracy of the annotation which is related to the complexity of the genome CDS structures and the degree to which the modelling procedure describes the genome. It is clear that integrating some comparative information into this approach improves the sensitivity of the method, particularly with respect to the overlapping regions that are missed by the SSA method. It is not always possible to run such a comparative analysis as there are not always additional sufficiently divergent sequences to run a comparative annotation with. However present comparative results indicate that where there is extra sequence information, the Phylo-HMM methodology will outperform the SSA. The SSA model is flexible enough to accommodate the differentiated nucleotide compositional signal that overlapping CDS provide to the HMM methodology. One aim of this work was to develop a method that maximizes the information that a single sequence can provide with respect to overlapping CDS, and this has been achieved to a large extent.

Insofar as a genome presented to this methodology violates the general assumptions or design specifications of the method, it will obviously not perform as well. The SSA method is specifically designed to annotate single stranded viral genomes and does not look for overlapping CDS in any opposing strand. In addition this method is not designed to deal with circular genomes although this could be accommodated with relative ease. This method does not identify introns readily and this is something which could possibly be addressed in further extensions. The model does not explicitly model instances of ribosomal slippage and transcriptional editing, but as the results on the Ebola genome indicate, sometimes this slippage is predicted as a CDS overlap and the main coding regions are still correctly annotated.

The Phylo-HMM framework for incorporating comparative information into the modelling procedure has been discussed and results indicate that adding evolutionary information does improve the annotation. An average increase of $6.73\%$ in the sensitivity measurements was observed for a $0.92\%$ cost in specificity when

...
applied to the HIV2 allignment of 14 sequences. This represents a significant improvement over the SSA method, and it is observable that the main improvement occurs in the overlapping regions where there is a 22% sensitivity boost. The authors are thus motivated to pursue this line of research further and intend to automate the parameterization of the mutational models and apply it to a series of ssRNA alignments. It is also interesting to consider whether incorporating some prior annotative knowledge (of which there is now plenty) into the process might help identify any hidden CDS regions. This could be achieved by placing restrictions on the Viterbi maximally likely search space, e.g. constraining the HMM to take a path consistent with a known CDS region in a given reading frame.

ACKNOWLEDGEMENTS

The authors would like to thank the members of our group who kindly read the manuscript including Gerton Lunter, Niall Mimouni and Saskia deGroot. Particular thanks to Gerton Lunter for suggestions which have improved the methodological description. We would also like to thank the anonymous referees for their helpful comments.

Conflict of Interest: none declared.

REFERENCES


