Sequence analysis

Comparative annotation of viral genomes with non-conserved gene structure

Saskia de Groot*, Thomas Mailund and Jotun Hein

Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG, UK

ABSTRACT

Motivation: Detecting genes in viral genomes is a complex task. Due to the biological necessity of them being constrained in length, RNA viruses in particular tend to code in overlapping reading frames. Since one amino acid is encoded by a triplet of nucleic acids, up to three genes may be coded for simultaneously in one direction. Conventional hidden Markov model (HMM)-based gene-finding algorithms may typically find it difficult to identify multiple coding regions, since in general their topologies do not allow for the presence of overlapping or nested genes. Comparative methods have therefore been restricted to likelihood ratio tests on potential regions as to being double or single coding, using the fact that the constrictions forced upon multiple-coding nucleotides will result in atypical sequence evolution. Exploiting these same constraints, we present an HMM based gene-finding program, which allows for coding in unidirectional nested and overlapping reading frames, to annotate two homologous aligned viral genomes. Our method does not insist on conserved gene structure between the two sequences, thus making it applicable for the pairwise comparison of more distantly related sequences.

Results: We apply our method to 15 pairwise alignments of six different HIV2 genomes. Given sufficient evolutionary distance between the two sequences, we achieve sensitivity of ∼84–89% and specificity of ∼97–99.9%. We additionally annotate three pairwise alignments of the more distantly related HIV1 and HIV2, as well as of two different hepatitis viruses, attaining results of ∼87% sensitivity and ∼98.5% specificity. We subsequently incorporate prior knowledge by 'knowing' the gene structure of one sequence and annotating the other conditional on it. Boosting accuracy close to perfect we demonstrate that conservation of gene structure on top of nucleotide sequence is a valuable source of information, especially in distantly related genomes.

Availability: The Java code is available from the authors.

Contact: degroot@stats.ox.ac.uk

1 INTRODUCTION

Due to their general constraint in sequence length, RNA viruses tend to compact coding information by using overlapping reading frames. This means that some parts of the viral genome

are coding for several proteins simultaneously, either in regions whose terminal points overlap or which are fully nested one in another. Since one amino acid is encoded for by a triplet of nucleotides, each locus potentially may be coding in up to three different contexts, thus being subject to multiple evolutionary constraints at a time. If a nucleotide is coding for two genes simultaneously, and is therefore part of two different codons, then a mutation of it might lead to a synonymous substitution in one codon but to a non-synonymous substitution in the other. This particular evolutionary behaviour, together with the topology of overlapping genes, will make it challenging for most general state of the art methods to successfully annotate full viral genomes.

Moreover, viruses often have undergone much evolution and gene structure might have changed significantly over time. For example, HIV1 and HIV2 have nine genes each, however, only eight of them are homologous. HIV1 has the additional *vpu* gene which is involved in viral budding and enhancing virion release from the cell. The *vpr* gene in HIV1 has the dual function of inducing cell cycle arrest and being in charge of nuclear import, whereas in HIV2 these two functions are split between the *vpr* gene and the additional *vpx* gene. Start and stop codons have also been shifted quite drastically, so that a state of the art comparative approach, which insists on totally conserved gene structure, would run into serious problems. Motivated by this, we introduce an HMM which overcomes these restrictions by allowing for overlapping genes, as well as evolved gene structure, and has separate evolutionary models for regions of different coding complexity.

The presence of overlapping reading frames in viral, prokaryotic and rarely even eukaryotic organisms is a well studied and established field. Ding *et al.* (1994), Walewski *et al.* (2001), Fukuda *et al.* (2003), and Makalowska *et al.* (2005), amongst many others, provide documentations of such examples. Pavesi *et al.* (1997), investigated the nucleotide composition of overlapping coding regions using information theory indices. Several other authors, such as Mizokami *et al.* (1997), Rogozin *et al.* (2002), and Hughes *et al.* (2005), have additionally analyzed the evolutionary constraints and selections that multiple coding regions are subjected to.

As far as bioinformatic methods for the analysis and detection of overlapping genes are concerned, there are several different approaches. One of the first articles is

*To whom correspondence should be addressed.

by Hein and Støvlbæk (1995), introducing a model describing the evolutionary process particular to multiple coding sequences. It may be used, amongst other things, to estimate evolutionary parameters, selection factors and the likelihood of an annotation. Pedersen and Jensen (2001) improved on this in accuracy by developing an exact evolutionary model for a region encoding two genes simultaneously. The context dependency of a particular nucleotide is, in contrast to Hein and Støvlbæk's work, non-stationary over time. This does, however, force the authors to use a computationally expensive Markov chain Monte Carlo (MCMC) method for the estimation of the evolutionary parameters. Firth and Brown (2005) developed a novel statistic to test whether a homologous region is double or single coding. In their very recent article, McCauley and Hein (2006) incorporated the codon biases of multiple coding regions into an HMM framework, which annotates single whole viral genomes. Furthermore, the authors demonstrated how to extend their model to a phylogenetic HMM in which they include the evolutionary information provided by a multiple alignment of homologous sequences, resulting in excellent annotation results.

Using HMMs for gene prediction is a well-studied area. Hobolth and Jensen (2005) introduce an HMM that allows comparative analysis of sequences related by a phylogenetic tree, without relying on a prior alignment. It is an extension to handle multiple sequences of Meyer and Durbin (2002)'s slightly more complex version, dealing with pairwise statistical alignment and gene finding, which adds the regions pertaining to splice sites and introns to the states of their HMM.

Approaching the problem of comparative gene finding within multiple coding viral genomes from an HMM point of view has been deemed a difficult and computationally expensive task. Prior comparative HMM methodologies for viruses have used conventional single coding methods to search through the genome on different reading frames. This, however, loses the information given to us by the particular evolutionary constraints a multiple coding region is under, since one is effectively treating every region as single or non-coding. Others have searched for a large concentration of highly degenerate amino acids (Pavesi, 2000) or used a simple likelihood ratio test to discern whether a region is single or double coding (Firth and Brown, 2005).

Building on the evolutionary model introduced by Hein and Støvlbæk (1995), we demonstrate how to extend it to an HMM for gene structure prediction of two aligned homologous viral genomes. Our HMM explicitly models all 64 possible multiple coding combinations in two sequences, there being eight in each. We thus allow for gene structure to have changed over time, which adds additional complexity to the method, differentiating it from most comparative gene finders.

We purposefully do not model gene length distribution, and including this could improve our annotation accuracy, at the cost of complicating our model slightly. However, for the time being we wish to evaluate the information provided by evolutionary behaviour alone, and any signal due to gene length distribution would threaten to overpower our results.

2 METHODS

2.1 Basic structure of our HMM

As usual, we will specify an HMM by five components: the set of states S, the matrix of transition probabilities $A = a_{ij}$, the emission alphabet Σ, the emission distribution e and the initial state distribution B. When in state i we have a certain probability e_i^c of emitting an element c from the alphabet Σ. In every state i, we may switch to another state j with probability a_{ij}. A path $\pi = (\pi)_K$ of visited states of length K is found by choosing the first state from the distribution B and following this by $K - 1$ state transitions according to A. This implies that the probability of observing a certain sequence $x = (x)_K$ together with a path π is given by

$$P(x, \pi | A, B, e) = B(\pi_1)e_{\pi_1}^{x_1} \prod_{k=2}^{K} e_{\pi_k}^{x_k} a_{\pi_{k-1}, \pi_k} \tag{1}$$

There are three unidirectional global reading frames, fixed before annotation of the sequence, which will be henceforth known as GRF1, GRF2 and GRF3. Each sequence may be coding for up to three genes simultaneously, and may thus be in one of the $2^3 = 8$ possible combinations of the three reading frames. Let us, for each sequence, visualize these states as the vertices of a unit cube (see Fig. 1). Since we are allowing for evolved gene structure, the state space S equals the cross-product of vertices of the two cubes, with $|S| = 64$.

We are given a gapped alignment of two homologous viral genomes. Since we are emitting pairs of nucleotides and must allow for gaps, our alphabet Σ will be over $\{A, C, G, T, -\} \times \{A, C, G, T, -\}$, where gaps in either sequence are treated as missing data. Every coding region starts with the start codon ATG and ends in one of the stop codons TAG, TAA or TGA. Generally, an HMM for gene finding in one reading frame would have a non-coding, a START, a coding and a STOP state, however, for the purpose of scanning sequences for genes we optimize this by introducing conditional transition probabilities. We may thus only cross from non-coding to coding in a certain reading frame if we have encountered a start codon in that reading frame and similarly transition from coding to non-coding conditional on finding a stop codon. Additionally, we set the silent start and end states to be non-
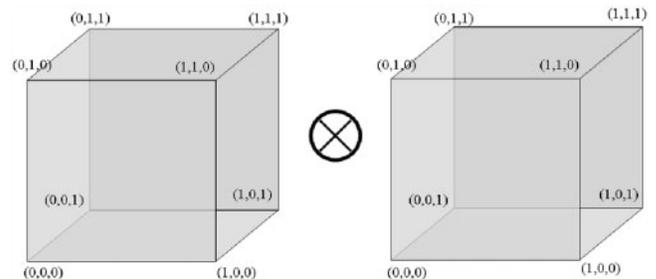
Fig. 1. The hypercube representing the 64 states, the two sequences can jointly be in. The vertices of the left cube represent the eight states the first sequence may be in. Here (0,0,0) is non-coding, whereas (1,1,1) is triple coding. Similarly, the right cube represents the states of the second sequence. Since we are allowing for a change in gene structure— i.e. we are not constraining the two sequences to be in the same state— they can be in any of the 8×8 combinations of the 'cross-product' of the two cubes.

coding in all three reading frames, thus ensuring to annotate only 'entire' genes as coding.

2.2 Transition probabilities

Each single sequence may be in one of the following eight states:

- (0,0,0) - non-coding
- (1,0,0) - coding in GRF1 only
- (0,1,0) - coding in GRF2 only
- (0,0,1) - coding in GRF3 only
- (1,1,0) - coding in GRF1 and GRF2
- (1,0,1) - coding in GRF1 and GRF3
- (0,1,1) - coding in GRF2 and GRF3
- (1,1,1) - coding in GRF1, GRF2 and GRF3

Since we are allowing for non-conserved gene structure, the pair of sequences may be in any of the 64 joint combinations of the above. When walking through our alignment of the two sequences, we only consider three different scenarios for entering a coding region in a particular reading frame (Fig. 2), conditional on finding a start codon in the respective sequence:

- Both sequences are non-coding in GRFx, we scan an aligned ATG in GRFx in both and both switch to coding—transition probability α.
- Both sequences are non-coding in GRFx, we scan an ATG in one sequence in GRFx but not in the other and that one switches to coding—transition probability β.
- One of the sequences is already coding in GRFx, we scan an ATG in the other in GRFx and it switches into coding as well—transition probability γ.

Regarding stop codons, if scanned with respect to a certain reading frame in which a sequence is coding, we switch into non-coding in that reading frame with probability 1. We do, in the above, make the assumption that if we are non-coding in both sequences and encounter an aligned ATG, then either both sequences switch to coding or both remain non-coding. If indeed this is an unfair assumption, we may easily adapt our model.

2.3 Emission probabilities

Several models of varying complexity have been devised to describe the evolutionary substitution process between two nucleotide sequences. We will be working with an extension of the Kimura two-parameter model (Kimura, 1980). The Kimura model assumes that transitions and
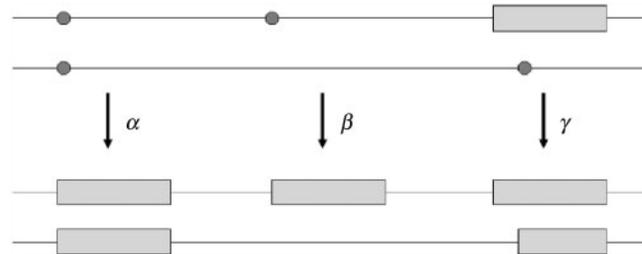
Fig. 2. The three possible scenarios for entering a coding region in a particular reading frame, with their respective probabilities of transition α, β, γ. An ATG is represented by a dot, a coding region in the particular reading frame by a box.

transversions occur at an instantaneous rate of g_{ts} and g_{tv}, respectively. If for two sequences, the evolutionary distance to the most recent common ancestor is $\tau/2$, we may write down the expected number of transitions and transversions per site as $a = g_{ts} \cdot \tau$ and $b = g_{tv} \cdot \tau$. We will be working solely with a and b and will thus not be able to separate out g_{ts}, g_{tv} and τ individually. Let P_{id}, P_{ts} and P_{tv} be the probabilities of, after time τ, at a certain locus observing an identity, transition and transversion, respectively. These are given by $\exp \mathbf{Q}\tau$ where $\mathbf{Q}$ is the instantaneous Kimura rate matrix:

$$P_{id}(a, b) = 1/4 \cdot (1 + \exp(-4b) + 2\exp(-2(a + b))) \qquad (2)$$
$$P_{ts}(a, b) = 1/4 \cdot (1 + \exp(-4b) - 2\exp(-2(a + b))) \qquad (3)$$
$$P_{tv}(a, b) = 1/2 \cdot (1 + \exp(-4b)) \qquad (4)$$

Since most amino acids are encoded by several different codons, Li *et al.* (1985) subsequently extended this idea by splitting each nucleotide position within a codon context into three different degeneracies. We may count the number of distinct amino acids arising when one alters each of the three nucleotide positions in each of the 16 potential contexts, and from this classify the nucleotides into three categories. A mutation of the position resulting in

- four times the same amino acid—Li denotes this as a site of degeneracy 4.
- two different amino acids, depending on whether a transition or transversion occurred 2:2.
- four different amino acids, regardless of the type of substitution 1:1:1:1.

We shorthand these as 4, 2 and 1. This approach brings some inherent problems with it, since not every site is classifiable as one of the above three degeneracies. For example, ATx codes for three isoleucines and one methionine and CGG and GGG are synonymous although one results from the other by a transversion. We will, for now, opt to restrict each degeneracy to one of the three above, realizing that this may be an unsatisfactorily inaccurate solution in the long run. For now, we classify 2:1:1 and 3:1 as 1:1:1:1, the latter due to the nature of ATG being a start codon.

We model the evolution of our sequences according to the Hein and Støvlbæk (1995) model, which is essentially an extension of the above to the overlapping reading frame context. When looking at a nucleotide in the ancestral sequence, for each reading frame we assign a certain state-dependent 'degeneracy-annotation' t to it, depending on its context. This will in a coding region in a particular reading frame be either of degeneracy 1, 2 or 4, equivalent to Li's notation, and for non-coding will always be designated as 0. Since we are considering overlapping reading frames, we thus obtain for each nucleotide in the ancestral sequence a certain state-dependent 'degeneracy-annotation-array' $t = [t1, t2, t3]$—an array consisting of the degeneracy annotation of a nucleotide for each of the three reading frames.

Using this degeneracy annotation, Hein and Støvlbæk incorporate the concept of selection factors into their framework: transitions and transversions occur according to the Kimura model, and non-synonymous substitutions get accepted by a factor f. Suppose a locus is of degeneracy [4,0,2], i.e. a locus is coding in GRF1 and GRF3 say for gene A and gene B, respectively. A change in nucleotide will result in a synonymous substitution in both reading frames if it is a transition, and in a non-synonymous one in GRF3 if it is a transversion. Thus our transition factor a remains as such, but we multiply our transversion factor b by the selection factor f_B for gene B.

So, let these factors for each nucleotide position i and each degeneracy-annotation array t be given by $F_i([t1, t2, t3])_{ts}$ and $F_i([t1, t2, t3])_{tv}$. We let these be dependent on f_1, f_2 and f_3, the selection factors for reading frame 1, 2 and 3. Assuming independence between

Table 1. Selection factors for each degeneracy-annotation array

	1:1:1:1	2:2	4	
1:1:1:1	$f_1f_2f_3, f_1f_2f_3$	$f_2f_3, f_1f_2f_3$	f_2f_3, f_2f_3	1:1:1:1
	$f_1f_2, f_1f_2f_3$	$f_2, f_1f_2f_3$	f_2, f_2f_3	2:2
	f_1f_2, f_1f_2	f_2, f_1f_2	f_2, f_2	4
2:2	$f_1f_3, f_1f_2f_3$	$f_3, f_1f_2f_3$	f_3, f_2f_3	1:1:1:1
	$f_1, f_1f_2f_3$	$1, f_1f_2f_3$	$1, f_2f_3$	2:2
	f_1, f_1f_2	$1, f_1f_2$	$1, f_2$	4
4	f_1f_3, f_1f_3	f_3, f_1f_3	f_3, f_3	1:1:1:1
	f_1, f_1f_3	$1, f_1f_3$	$1, f_3$	2:2
	f_1, f_1	$1, f_1$	$1, 1$	4

The selection factors denoted as $F([t1, t2, t3])_{ts}$, $F([t1, t2, t3])_{tv}$ which are to be multiplied onto the basic transition and transversion parameters a and b. The top axis refers to the first, the left to the second and the right to the third global reading frame. Note that a non-coding site will be treated the same as a site of degeneracy 4. We are assuming independence of genes since otherwise $f_1 \cdot f_2$ would be replaced by f_{12}.

genes, the probability of a mutation occurring gets multiplied up by the selection factor of each reading frame that it causes a non-synonymous change in. Then the probabilities of observing at a site of degeneracy $[t1, t2, t3]$ an identity, transition and transversion after time τ are given by

$$P_{id}\left(\tilde{a}, \tilde{b}\right) = 1/4 \cdot (1 + \exp(-4\tilde{b}) + 2\exp(-2(\tilde{a} + \tilde{b}))) \quad (5)$$

$$P_{ts}\left(\tilde{a}, \tilde{b}\right) = 1/4 \cdot (1 + \exp(-4\tilde{b}) - 2\exp(-2(\tilde{a} + \tilde{b}))) \quad (6)$$

$$P_{tv}\left(\tilde{a}, \tilde{b}\right) = 1/2 \cdot (1 + \exp(-4\tilde{b})) \quad (7)$$

where

$$\tilde{a} = a \cdot F([t1, t2, t3])_{ts} \quad (8)$$

$$\tilde{b} = b \cdot F([t1, t2, t3])_{tv} \quad (9)$$

with F as given in Table 1. Note that our evolutionary model requires the two-sided coding context of each nucleotide in the ancestral sequence to be able to ascertain the degeneracy annotation. Since we are not modelling the ancestral sequence composition, but more the evolution to the second sequence conditional on the composition of the first, we are still working in a Markovian framework and all general theorems hold. Although the coding context will depend on which sequence is chosen as ancestral, in fact the vast majority of contexts remain identical throughout evolution, so that in all pairwise comparisons our results differ only minimally with our choice of ancestor.

Additional consideration needs to be given to regions where the gene structure has changed. In regions where gene structure differs, we cannot hope to discern any useful signal unless the structural change has occurred very recently—indeed our method picks up structural change merely by finding start and stop codons which are compatible with a conserved region. Since we wish our model to be time reversible to the greatest extent, we therefore decide to model the evolution of regions coding in only one sequence as unconstrained, i.e. equivalent to non-coding.

2.4 Parameter estimation

Having devised our model, we want to apply it to annotate two aligned genomes. Our model parameters are given by $\Theta = [\alpha, \beta, \gamma, a, b, f_1, f_2, f_3]$, so we wish to find those which maximize the likelihood of our data. In the case of our parameters being free, we could simply use the Baum–Welch algorithm for this, however, our scenario is not quite that simple.

In the case of the transition probabilities, we have three parameters: α, β and γ. Thus, we do not wish to work out the expected number A_{ij} of times we transitioned from state i to state j, but instead the expected number of times that a transition of type α, β and γ occurred. For the case of α, say, just group all expected transitions A_{ij} of type α together and call this number E_α. We also work out the expected number of times that the transition α was not made and call this $E_{1-\alpha}$. Remember, that since we have three different types of transition, we may not simply look at the total number of transitions made. Then our maximum likelihood estimator for α is given by

$$\hat{\alpha} = \frac{E_\alpha}{E_\alpha + E_{1-\alpha}} \quad (10)$$

and similarly so for β and γ.

When we consider the emission probabilities, we remember that our emissions fall into several different degeneracies according to their nucleotide context. We calculate, using the forward–backward probabilities, for each degeneracy the expected number of times an identity, transition and transversion is used. For a site of degeneracy $t = [t1, t2, t3]$, let this be $x_{id,t}$, $x_{ts,t}$ and $x_{tv,t}$, respectively. Since $P_{id,t}$, $P_{ts,t}$ and $P_{tv,t}$ were the probabilities for a site of degeneracy t of an identity, transition or transversion occurring [see Equations (5–7)], we may rewrite the emission term of the log likelihood as follows

$$\sum_i \sum_t x_{id,t} \log P_{id,t} + x_{ts,t} \log P_{ts,t} + x_{tv,t} \log P_{tv,t}$$

For this function of the five emission parameters a, b, f_1, f_2 and f_3, we now find the maximum likelihood estimates using the Newton–Raphson iteration method and repeat the estimation step. Once the likelihood has converged, we use the Viterbi algorithm to find the most likely state annotation of the sequence alignment (Durbin *et al.*, 1998).

2.5 Sensitivity and specificity scoring

When evaluating the accuracy of our annotation, we must think of a prudent way to define a sensitivity and specificity score. An annotation correct in one reading frame and false in another is, using normal methods, not easily classifiable. We therefore need a measure which draws the complexity of potentially coding in up to three reading frames into account.

For the sake of direct comparison, we adopt the method introduced by McCauley and Hein (2006). As true positives we take the sum $\sum_i C^+(x_i)$ where, x_i is the ith nucleotide and $C^+(x_i)$ is the number of reading frames it is coding in. Similarly, we define the true negatives to be $\sum_i C^-(x_i)$ where, $C^-(x_i)$ is the number of reading frames the nucleotide is not coding in. Then we may as usual define

- Sensitivity $= \dfrac{(TP - FN)}{TP}$

- Specificity $= \dfrac{(TN - FP)}{TN}$

where TP, FP, TN and FN are true and false positives and negatives, respectively. Since we are annotating both sequences simultaneously, we give our sensitivity and specificity scores as an average over both sequences.

Table 2. Sensitivity, specificity and parameter estimates for HIV2 versus HIV2

Sequences	Sensitivity	Specificity	a	b	f_1	f_2	f_3
J04542 - U27200	0.9000	0.9990	0.283	0.114	0.360	0.250	0.413
J04542 - L36874	0.8312	0.9998	0.253	0.088	0.279	0.283	0.583
M15390 - U27200	0.8985	0.9994	0.260	0.115	0.243	0.339	0.492
M15390 - L36874	0.8316	0.9758	0.226	0.082	0.388	0.469	0.294
M15390 - J04542	0.8315	0.9911	0.155	0.028	0.277	0.428	0.665
U27200 - L36874	0.7973	0.9756	0.180	0.045	0.456	0.152	0.439
M30502 - U27200	0.8759	0.9732	0.266	0.107	0.422	0.516	0.271
M30502 - J04542	0.7893	0.9825	0.081	0.013	0.687	0.368	0.381
M30502 - M15390	0.8447	0.9654	0.148	0.031	0.248	0.441	0.803
M30502 - L36874	0.8400	0.9618	0.256	0.082	0.340	0.401	0.481
U27200 - D00835	0.8919	0.9748	0.263	0.115	0.399	0.424	0.240
M15390 - D00835	0.5050	1.0000	0.090	0.015	0.392	1.091	0.244
J04542 - D00835	0.8553	1.0000	0.147	0.027	0.255	0.495	0.685
L36874 - D00835	0.8518	0.9706	0.240	0.090	0.267	0.303	0.471
M30502 - D00835	0.8158	0.9639	0.144	0.032	0.497	0.597	0.238

Sensitivity and specificity comparisons on the fifteen pairwise genome annotations of six different HIV2 strains. To the right are given the parameter estimates of the transition/transversion rates a and b as well as the selection factors f_1, f_2 and f_3 for the three different reading frames. Note that the same genes might be in different global reading frames in the various pairwise alignments so one cannot expect the predictions to be equivalent within one column. A graphical representation of this table is given in Figure 4.

3 RESULTS

3.1 Simulated data

Initially we wish to test our method on simulated data. We took several HIV genomes from GenBank and let them evolve with varying evolutionary parameters ranging from an evolutionary distance of $a + 2b = 0.1$–3.0. We then annotated the resulting alignment. Everything above a distance of ~ 0.2 was estimated to very high accuracy generally reaching a sensitivity and specificity of about 99%. When dealing with more closely related descendant sequences, we started to encounter severe problems below the 0.15 mark and sensitivities plummeted down to 70%. We generally estimated the transition and transversion rates a and b to $\sim 5\%$ of their true value, regardless of evolutionary distance. Our parameter estimates of the selection factors—tested between 0.1 and 1.0—were good and generally around ± 0.035 of their true value. However, for more closely related sequences the quality of estimation for selection factors was much more volatile going up to ± 0.2 wrong in some cases. Also the loss in sensitivity was nearly always due to us missing out the short intronic *rev* gene, even in sequences far apart, which brings up the question whether a short region can ever provide a strong enough signal to be picked up on by our method as coding. Specificity loss was generally due to a double coding region being designated as triple coding in the presence of an additional short open reading frame.
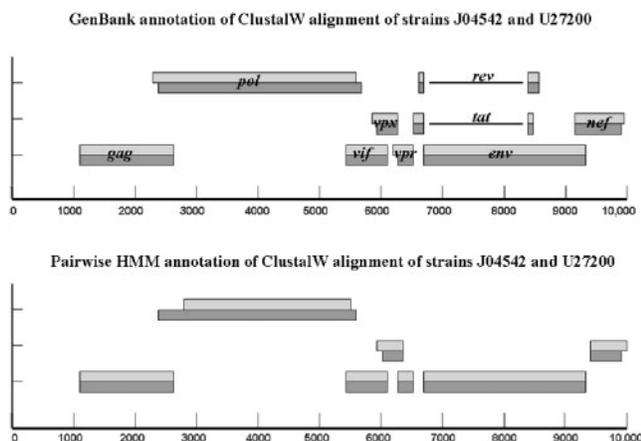
3.2 Data preparation

We downloaded pairs of viral sequences from the GenBank database and used CLUSTALW (Thompson *et al.*, 1994) to obtain a pairwise gapped alignment. We heavily rely on gaps within coding regions occurring in triplets. After the CLUSTALW alignment, we therefore manually adjusted the sequences for this, which is generally a trivial exercise. We obtained the seed parameters for the expectation-maximization (EM) algorithm by marking every open reading frame above 200 nt as coding and subsequently calculated from this the maximum likelihood estimates of our parameters α, β, γ, a, b, f_1, f_2 and f_3.

3.3 Pairs of HIV2

We performed a pairwise comparison on the fifteen different combinations of the six HIV2 strands with GenBank accession number J04542, M15390, U27200, L36874, M30502 and D00835. The results are illustrated in Table 2. We took a change of < 1 in log likelihood as an indication of completed convergence and usually the EM algorithm converged sufficiently after about three iterations. A particular pairwise annotation of U27200 and J04542 is shown in Figure 3, with the respective GenBank annotation above it. As one can see, our program misses out on the two very short intronic genes and misannotates the *pol* gene due to ribosomal slippage having occurred in the J04542 strand. It also starts annotating the *nef* gene ~ 200 bp too late, presumably due to lack of conservation. Otherwise, the genes are correctly identified—even where the start and stop codons have shifted—and we achieve a sensitivity and specificity of around 89.7 and 99.9%. Concentrating on purely the overlapping regions, we achieve a sensitivity of $\sim 68\%$. This in particular distinguishes our method from other comparative approaches, which would not be able to discern multiple coding regions as such, see Section 7.7.

Table 2 and Figure 4 show that we need sequences to have an appropriate evolutionary distance to obtain any reasonable results, as our simulations already suggested. Generally, our standard errors were around 0.03 for transition and transversion rates and between 0.03 and 0.1 for the selection factors. The error estimate for the selection factor not pertaining to either the *gag* or *pol* reading frame was unsurprisingly

GenBank annotation of ClustalW alignment of strains J04542 and U27200

Pairwise HMM annotation of ClustalW alignment of strains J04542 and U27200

Fig. 3. The annotation of HIV2 strands of U27200 and J04542. Above is the GenBank annotation and below the prediction of our program. Each bar shows the genes in one sequence, with intronic regions being marked by single lines. Where a pair of bars does not overlap this indicates the change in gene structure via a shift in start or stop codon.
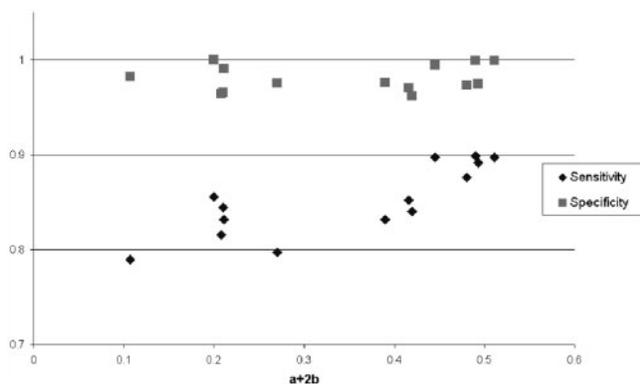
Fig. 4. A graph representation of Table 2 (without the M15390-D00835 comparison), with the evolutionary distance $a + 2b$ along the x-axis, and sensitivity and specificity along the y-axis. We can see that with growing evolutionary distance our predictions tend to become better. Ribosomal slippage occurring in some strands also accounts for some of the fluctuation in prediction accuracy.

consistently slightly higher than the other two, due to the length of these genes.

3.4 HIV1 versus HIV2

We also ran the program on three sequence alignments of HIV1 and HIV2 genomes. This was naturally a much more challenging exercise, since the two sequences are more divergent and thus the gene structure has changed substantially in some areas. Moreover, presumably due to the large evolutionary distance between these two different virus strains, CLUSTALW gives a very inaccurate alignment. Indeed, an accurate *ab initio* alignment of these sequences is currently not feasible. We therefore, for now, use the program GenAl introduced by Hein and Støvlbæk (1994, 1996), which combines DNA and protein

Table 3. Individual posterior parameter estimates for HIV1 versus HIV2

| Gene	a	b	$a+2b$	a/b	f
non	0.37	0.21	0.79	1.72	N/A
gag	0.38	0.26	0.9	1.52	0.46
pol	0.35	0.24	0.83	1.44	0.46
vif	0.38	0.28	0.94	1.36	0.82
vpr	0.78	0.47	1.72	1.67	0.29
tat	0.24	0.21	0.66	1.12	0.89
rev	N/A	N/A	N/A	N/A	N/A
env	0.46	0.45	1.36	1.01	0.49
nef	0.34	0.27	0.88	1.28	0.52
non	0.33	0.25	0.83	1.33	N/A

The different individual maximum likelihood estimates of the transition-transversion rates a and b, their ratio a/b, the evolutionary distance $a + 2b$ and the selection factor f for each region of HIV1-HIV2 comparison.

alignment, in particular for genomes with overlapping reading frames. As an input GenAl has both individual sequences and a list of coding regions, taken from GenBank. It subsequently optimizes both the DNA and the protein alignment simultaneously, whilst allowing for the presence of multiple coding regions. This will naturally be a problem when doing *de novo* gene annotation, since GenAl requires a list of coding regions for its alignment, however, for the sake of our purpose it must suffice for the moment.

Nonetheless, we encounter some difficulties since homologous genes have undergone a frameshift over time, due to indels of length not multiples of three. Our program is dependent on a homologous region being coding in the same reading frame. Additionally, the presence of non-triplets of gaps within a gene will generally result in the premature presence of a stop codon. We may minimize this problem, however, by manually adding single 'fake' pairs of gaps to both sequences thus bringing the sequence regions into the correct global reading frame again without changing the actual alignment.

Altogether, when comparing HIV1 and HIV2 we achieve an average sensitivity of $\sim 80\%$ and a specificity of $\sim 98.5\%$. Standard errors were smaller than in the HIV2-HIV2 comparison, generally around 0.035 but again slightly higher for the one selection factor which was not used for either *gag* or *pol*. When obtaining our HMM annotation several features stand out. We are encouraged to see that genes with shifted start and stop codons generally get annotated correctly. Using a pairwise comparative approach, we cannot expect the non-homologous *vpx* and *vpu* genes to be annotated. Also, the very short *rev* and *tat* genes are very difficult to pick up on.

Clearly within coding regions and non-coding regions evolutionary rates will generally differ substantially, however, there will also be intergenic differences due to distinct selection factors and intragenic differences due to slow and fast evolving regions. We fixed the GenBank annotation and estimated the evolutionary parameters for each individual region. Our sensitivity problems mainly boil down to our annotation missing out on the *vif* and a few hundred nucleotides of the
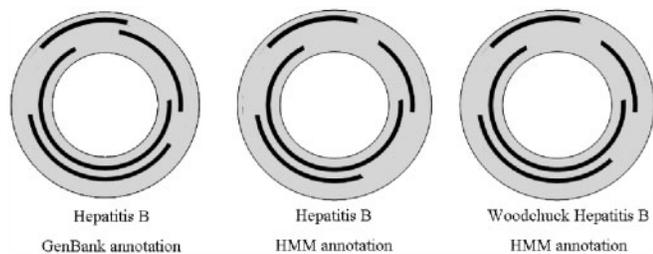
Fig. 5. The circular Hepatitis B genome. To the left the GenBank annotation, which is nearly the same for both the Hepatitis B NC003977 and the Hepatitis B Woodchuck J02442 strands. To the right are the annotations for each strand predicted by our method. The dissimilarities in annotation arise due to our method choosing unaligned start codons further downstream than the true one.

env gene. Looking at Table 3 we can see the individual parameter estimates for these regions. *Gag* and *pol* will highly dominate the maximum likelihood estimates for the evolutionary parameters, due to their length. *Vif* has average evolutionary rates but a very high selection factor of 0.8, whereas *env* has an average selection factor but very much higher transition and transversion rate estimates than *gag* and *pol*. Their non-conformity may be an indication as to why these regions account most for our loss of sensitivity and also suggest that it is problematic to assume constant transition and transversion rates along the genome.

3.5 Hepatitis B virus

We also ran our method on an alignment of Hepatitis B strand NC003977 and Woodchuck Hepatitis B strand J02442, as these are known to contain large sections of overlapping coding regions. Due to the circular nature of the Hepatitis B genome, we adjoined two copies of each strand to one another and aligned these using CLUSTALW. We subsequently cut the alignment in the only non-coding region of the Woodchuck Hepatitis strand and discarded the repeated bits 100 nt to the left and right of the cut. Seed parameters were obtained as before, and from this an annotation was generated using our EM algorithm.

The evolutionary distance was estimated at $a + 2b = 0.96$ with $a = 0.38$ and $b = 0.29$, thus being comparable to HIV1 and HIV2 in phylogenetic proximity, albeit with more conserved gene structure. We managed to recover ∼83% of the overlapping regions, suggesting that our evolutionary model, though adequate, is not entirely satisfactory in its description of multiple coding regions. Nonetheless, we achieve an overall sensitivity and specificity of 87.4 and 98.8%, respectively—an encouraging result, considering the complexity of the Hepatitis B virus. A picture of the annotation is given in Figure 5.

We also ran our method on the reverse complement of the Hepatitis B alignment, to test whether the presence of a reverse encoded gene could cause a false positive prediction in the forward strand. Two short ORFs, overlapping conserved coding regions in the complement strand, were marked as coding, resulting in a specificity of 98.5%.
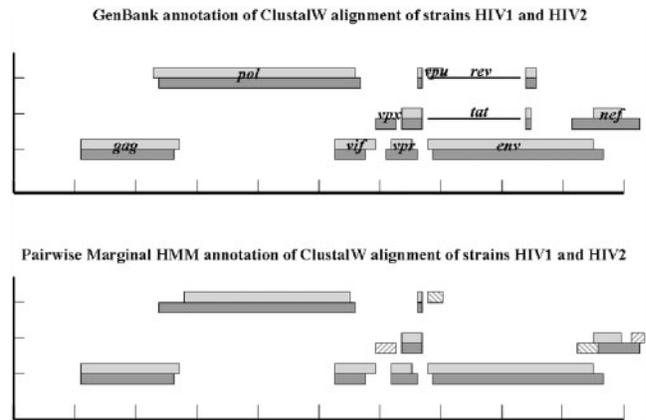
Fig. 6. Above is the GenBank annotation and below the two marginal annotations of genomes K02013(HIV1) and M30502(HIV2). The dashed areas are the mispredicted areas in one of the marginal annotations, with left and right slanted dashes being annotated only when conditioning on HIV1 and HIV2, respectively.

3.6 Incorporating prior knowledge

We have developed a methodology for annotating two unknown homologous viruses. Many virus families, however, are reasonably well studied, HIV in particular. When a new virus is sequenced, it would be of far greater use to annotate it using our existent knowledge of similar genomes. For argument's sake, suppose HIV2 had just been discovered—a virus belonging to the same family as HIV1 but structurally slightly different. Annotating the HIV2 virus *ab initio* would be throwing away a lot of prior knowledge. On the other hand, state of the art comparative annotation would be assuming common gene structure, which does not hold. We will therefore adapt our above methodology to tackle this particular problem.

Assuming HIV1 is known and well studied, we most likely will obtain a highly reliable annotation off GenBank. In our above representation, the annotation of HIV1 will happen on one cube and the annotation of HIV2 on the other (see Fig. 1). If we know the annotation of HIV1, then this is equivalent to being able to fix the state path along that cube. We then want to find the most likely state path for HIV2 *given* HIV1. This is easily incorporated into our above methodology by weighting the Forward, Backward and Viterbi probabilities accordingly. We need to weight them, as opposed to deterministically restrict them, since we must draw into account the possibility of the GenBank annotation being inaccurate and us fixing a path invalid under our model. Let $s = (s)_i$ be the true state path through the annotation, as annotated in GenBank. Biasing our annotation translates into multiplying the Viterbi probability of being in state k at position i by a factor $|\mathbb{I}_{k=s_i} - \epsilon|$, where, $\mathbb{I}$ is the indicator function and ϵ may be chosen to be as strong a weight as desired. The weighted forward and backward probabilities are calculated accordingly.

To test our approach, we use the same genomes as in Section 3.4 and in our Viterbi annotation fix the annotation of HIV1. For sake of comparison, we subsequently annotate HIV1 given HIV2. The results are close to perfect, achieving

Table 4. Comparison to other methods

Method	Overall sensitivity	Overlapping sensitivity	Overall specificity
Firth and Brown	1.000	1.000	0.932
GeneMark.hmm	0.382	0.137	1.000
GLIMMER	0.589	0.286	0.974
McCauley and Hein Phylo	0.897	0.847	0.982
de Groot *et al.*	0.874	0.830	0.988

Sensitivity and specificity results of several methods on the Hepatitis B strand NC003977. Separate attention is given to the accuracy when restricted to overlapping regions. For direct comparison, we disregard any false positive predictions which occur on the reverse complement strand.

sensitivity of 96 and 99%, respectively, and specificity of 99.7%. Similarly, on the Hepatitis alignment we achieve 100% sensitivity and 100% specificity for Hepatitis B and 100% sensitivity and 94% specificity for the Woodchuck Hepatitis, respectively conditional on the other.

Although parameter estimates differ slightly, the final Viterbi annotation shown in Figure 6 of the two sequences is close to identical, apart from naturally both times the non-homologous genes in the other sequence not being picked up on. The only remarkable thing is the beginning of the *nef* gene in the HIV2 strand not being annotated, when conditioning on the HIV1 strand. This is presumably due to lack of conservation in that area, as noted before. Also, the estimated state transition probabilities in both marginal joint annotations are very close. The improvement on annotation accuracy is dramatic, and demonstrates the amount of knowledge still maintained between the two sequences, due to their structural similarity. Basically, the only features we still miss out on are the ones our model is incapable of capturing: non-homologous genes, introns and ribosomal slippage. Although on a nucleotide level both sequences differ quite substantially with only ~50% sequence similarity, the structural conservation over time provides us with enough information to annotate the homologous regions in the 'unknown' strand highly successfully.

3.7 Comparison to other methods

When comparing our results to other methods (see Table 4), several aspects must be drawn into account. Most available comparative gene finders, such as SLAM, TWAIN and TWINSCAN (Alexandersson *et al.*, 2003; Korf *et al.*, 2001; Majoros *et al.*, 2005), are configured towards eukaryotes, and thus not applicable to viruses. GLIMMER by Salzberg *et al.* (1998) is a gene finder designed towards microbial genomes, which results in 58.9% sensitivity, and specificity of 97.4%, recovering merely 29.6% of overlapping regions, even though it is designed to accommodate for these. Similarly, GeneMark.hmm (Lukashin and Borodovsky, 1998)—used by Mills *et al.* (2003) to create the VIOLIN database—which achieves comparable results to ours on the HIV virus, runs into problems when annotating the Hepatitis B genome, leading to an overall sensitivity of merely 38%, with only 14% of overlapping regions being annotated as such.

Within approaches specifically designed towards multiple coding regions, McCauley and Hein (2006)'s single sequence method achieves similar results to ours. Their signal is purely taken from codon bias and gene length distribution, whereas ours is solely from comparative information. When extended to a phylogenetic model, accuracy is naturally boosted slightly higher, though it would be interesting to see how their method performed when applied to genomes with non-conserved gene structure. Still, as shown in Table 4 our performance is highly comparable even to the phylogenetic method, especially considering that our runtime is several orders of magnitude smaller.

Firth and Brown (2005, 2006) describe another comparative method capable of annotating multiple coding regions within a genome. Their approach is, however, more designed towards the detection of a novel overlapping gene, given a prior annotation. It would test the hypothesis of a query region being double as opposed to single coding. When used *ab initio* on an unannotated genome, they may only test whether a region is single as opposed to non-coding. This results under similar assumptions of minimal ORF length, in the annotation of many false overlapping genes in both reading directions and a comparably high false positive rate.

4 DISCUSSION

We have introduced a novel HMM approach for annotating two homologous genomes containing overlapping reading frames. Most importantly, our model is not restricted to conserved gene structure—a feature not realized in similar methods, since they generally insist on aligned start and stop codons (Hobolth and Jensen, 2005; Meyer and Durbin, 2002). Albeit just using evolutionary information and disregarding actual sequence composition—such as codon usage and CG richness—we achieve encouraging results. We correctly identify the 'normal' genes up to a very high level of accuracy, even when there has been a shift in start or stop codon over time. On homologous sequences of sufficient evolutionary distance, we expect a sensitivity of around 83−89% (depending on whether ribosomal slippage has occurred) and a specificity of 97−99.9%. On the non-homologous HIV1 and HIV2 comparison, we still keep a sensitivity of around 80% and a specificity of around 98.5%. On the highly complex Hepatitis B virus, we achieve a sensitivity and specificity of 88 and 99% respectively, recovering 83% of overlapping regions. We are thus highly competitive towards other state-of-the-art methods.

Our quality of prediction is highly dependent on sequences not being evolutionarily too close together, in which case our program finds it hard to pick up conservation due to functionality as opposed to mere phylogenetic proximity. As our simulation results demonstrate, this however is to be expected. Finally, we demonstrate how to annotate one sequence knowing another, given that they are structurally related, though not identical. We achieve close to perfect accuracy when annotating one sequence conditional on the other, both for HIV and for Hepatitis, demonstrating the power of information contained in gene structure conservation.

Since our program cannot deal with ribosomal slippage or the presence of introns, there is an inherent amount of failure in our method. We have theoretically extended our HMM topology to allow for introns, however, the resulting blow-up in state space to a total of 15625 states, and the resulting computational demands make this an infeasible option. It is also debatable, whether the signal provided would be strong enough to detect introns. Similarly, a simple state space extension could allow us to deal with the presence of reverse reading frames, but again complexity issues initially prohibit this.

Arguably, another drawback is our modelling of selection factors. Our assumption of independence, as shown by Hein and Støvlbæk (1995) appears to be a reasonable if not ideal one. In contrast, our premise of 'one reading frame—one selection factor'' is not biologically justifiable, and brings problems along with it, especially when looking at more diverse sequences. Ideally one would like every gene to have its own selection factor, drawn from a set of n strengths. However, this would result in an n^3-fold increase in state space and is thus, again, not a practical option. In contrast, constraining selection to be constant along the genome appears to drastically worsen our results. As several articles show, the smaller genes, such as *tat*, *rev*, *nef* and *vif* contain several sites under positive selection (see de Oliveira *et al.*, 2004; de Zanotto *et al.*, 1999; Yang and Swanson, 2002). The selection factor estimates in our model will be highly dominated by the longer *gag* and *pol* genes, which are believed to be under strong purifying selection (Seibert *et al.*, 1995). Thus allowing for reading frame specific as opposed to constant selection does not solve, but certainly greatly diminishes the problem of shorter genes getting over-powered in the parameter estimation. One way of modelling heterogenous selection along the genome, would be to introduce several auto-correlated selection strength classes and let, at each site, selection be chosen from one of them—a reasonably easy adaptation to our model.

The fact that our method is so sensitive to alignments, in the case of gaps within coding regions not occurring in triplets, is also undesirable. The complexity of our model implies that a gap singlet in one sequence can throw the entire annotation off balance. Additionally, we are in great need of an aligner for sequences far apart, such as HIV1 and HIV2. Our use of the alignment provided by GenAl is far from ideal, and we are currently concentrating on incorporating simultaneous alignment into our methodology as well as summing over all possible alignments. This would make the method readily available for *ab initio* comparative gene annotation with or without prior knowledge, in particular for more distantly related homologous genomes with non-conserved gene structure.

Including codon bias would be a natural extension to our model and could be accommodated by applying a method similar to the one described in McCauley and Hein (2006). This would basically involve turning the ancestral sequence into a second order Markov chain, where for each state, given the prior 2 nt, the following one is drawn from a multinomial distribution. With this method, however, one automatically assumes gene length to be geometrically distributed, and it is debatable how much this may overpower any evolutionary signal and result in little more than a long open reading frame (ORF) scanner.

We naturally would also like to extend our method to multiple sequences, since especially with viral genomes such data is readily at hand. However a blow-up in state space, and ambiguity as to how to define the transition parameters render this a far from trivial problem. Future work will concentrate on making an alignment independent method as well as improving our evolutionary model for overlapping regions and thus investigating the evolutionary pressures that underly these.

ACKNOWLEDGEMENTS

REFERENCES

Alexandersson,M. *et al.* (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.*, **13**, 496–502.

de Oliveira,T. *et al.* (2004) Mapping sites of positive selection and amino acid diversification in the HIV genome. *Genetics*, **167**, 1047–1058.

de Zanotto,P. *et al.* (1999) Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics*, **153**, 1077–1089.

Ding,S.W. *et al.* (1994) New overlapping gene encoded by the cucumber mosaic virus genome. *Virology*, **198**, 593–601.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press.

Firth,A.E. and Brown,C.M. (2005) Detecting overlapping coding sequences with pairwise alignments. *Bioinformatics*, **21**, 282–292.

Firth,A.E. and Brown,C.M. (2006) Detecting overlapping coding sequences in virus genomes. *BMC Bioinformatics*, **7**.

Fukuda,Y. *et al.* (2003) On dynamics of overlapping genes in bacterial genomes. *Gene*, **323**, 181–187.

Hein,J. and Støvlbæk,J. (1994) Genomic alignment. *J. Mol. Evol.*, **38**, 310–316.

Hein,J. and Støvlbæk,J. (1995) A maximum-likelihood approach to analyzing non-overlapping and overlapping reading frame. *J. Mol. Evol.*, **40**, 181–189.

Hein,J. and Støvlbæk,J. (1996) Combined DNA and protein alignment. *Meth. Enzymol.*, **266**, 402–418.

Hobolth,A. and Jensen,J.L. (2005) Applications of hidden Markov models for characterization of homologous DNA sequences with a common gene. *J. Comput. Biol.*, **12**, 186–203.

Hughes,A.L. and Hughes,M.A.K. (2005) Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Res.*, **113**, 81–88.

Johnson,Z.I. and Chisholm,S. (2006) Properties of overlapping genes are conserved across microbial genomes. *Genome Res.*, **14**, 2268–2272.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In: Munro H.N. (ed), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Korf,I. *et al.* (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.

Li,W.S. (1993) Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J. Mol. Evol* **36**, 96–99.

Li,W.S. *et al.* (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174.

Lukashin,A. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Majoros,W.H. *et al.* (2005) Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics*, **21**, 1782–1788.

Makalowska,I. *et al.* (2005) Overlapping genes in vertebrate genomes. *Comput. Biol. and Chem.*, **29**, 112.

McCauley,S. and Hein,J. (2006) Using HMMs and observed evolution to annotate viral genomes. *Bioinformatics*, Advance Access published online on April 13, 2006.

Meyer,I.M. and Durbin,R. (2002) Comparative ab initio prediction of gene structure using pair HMMs. *Bioinformatics*, **18**, 1309–1318.

Mills,R. *et al.* (2003) Improving gene annotation of complete viral genomes. *Nucleic Acids Res.*, **31**, 7041–7055.

Mizokami,M. *et al.* (1997) Constrained evolution with respect to gene overlap of Hepatitis B Virus. *J. Mol. Evol.*, **44**, 83–90.

Pavesi,A. (2000) Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. *J. Mol. Evol.*, **50**, 284–295.

Pavesi,A. *et al.* (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J. Mol. Evol.*, **44**, 625–631.

Pedersen,J.S. and Hein,J. (2003) Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, **19**, 219–227.

Pedersen,A.M. and Jensen,J.L. (2001) A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, **18**, 763–776.

Rogozin,I. *et al.* (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, **18**, 228–232.

Salzberg,S.L. *et al.*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.

Seibert,S. *et al.* (1995) Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.*, **12**, 803–813.

Thompson,J.D. *et al.* (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Yang,Z.H. and Swanson,W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.*, **19**, 49–57.

Walewski,J.L. *et al.* (2001) Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. *RNA*, **7**, 710–721.

All data used is publicly released on the GenBank database, see http://www.ncbi.nlm.nih.gov/

ClustalW Software can be found on the web at http://www.ebi.ac.uk/clustalw/