



Genome Annotation and Selectional  
Analysis of Viral Evolution

Saskia de Groot

Under Supervision of Prof. Jotun Hein

A thesis submitted for the degree of

*Doctor of Philosophy*

2007

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
1.1	Abstract . . . . .	6
1.2	Literature Review . . . . .	8
1.2.1	Biological Background . . . . .	8
1.2.2	Sequence Analysis of Overlapping Coding Regions . . .	13
1.2.3	Investigating Selection Acting on Overlapping Regions	16
1.2.4	Modelling Selection on Overlapping Regions . . . . .	22
1.2.5	Gene Annotation in General . . . . .	26
1.2.6	Gene Finding in Overlapping Reading Frames . . . . .	30
1.2.7	Decreasing Uncertainty when Using Alignments . . . . .	36
<b>2</b>	<b>Annotation of Viruses with Non-Conserved Gene Structure</b>	<b>38</b>
2.1	Abstract . . . . .	39
2.1.1	Motivation . . . . .	39
2.1.2	Results . . . . .	39
2.2	Introduction . . . . .	40
2.3	Methods . . . . .	42
2.3.1	Basic Structure of our HMM . . . . .	42

2.3.2	Transition Probabilities . . . . .	43
2.3.3	Emission Probabilities . . . . .	49
2.3.4	Parameter Estimation . . . . .	56
2.3.5	Sensitivity and Specificity Scoring . . . . .	60
2.4	Results . . . . .	60
2.4.1	Simulated Data . . . . .	60
2.4.2	Data Preparation . . . . .	61
2.4.3	Pairs of HIV2 . . . . .	62
2.4.4	HIV1 vs. HIV2 . . . . .	68
2.4.5	Hepatitis B Virus . . . . .	71
2.4.6	Incorporating Prior Knowledge . . . . .	72
2.4.7	Comparison to Other Methods . . . . .	74
<b>3</b>	<b>Annotation of Selection Strengths in Viral Genomes</b>	<b>78</b>
3.1	Abstract . . . . .	79
3.1.1	Motivation . . . . .	79
3.1.2	Results . . . . .	79
3.2	Introduction . . . . .	80
3.3	Methods . . . . .	82
3.3.1	Basic Structure of our Model . . . . .	82
3.3.2	Model Parameters . . . . .	85
3.4	Results . . . . .	90
3.4.1	Simulated Data . . . . .	90
3.4.2	HIV2 . . . . .	93
3.4.3	Hepatitis B . . . . .	100

<b>4</b>	<b>Investigating Selection: A Statistical Alignment Approach</b>	<b>106</b>
4.1	Abstract . . . . .	107
4.1.1	Motivation . . . . .	107
4.1.2	Results . . . . .	107
4.2	Introduction . . . . .	108
4.3	Methods . . . . .	110
4.3.1	Outline . . . . .	110
4.3.2	Substitution model . . . . .	112
4.3.3	Alignment model . . . . .	116
4.3.4	Full model . . . . .	118
4.3.5	Extension to Multiple Sequences . . . . .	119
4.4	Results . . . . .	120
4.4.1	Simulation . . . . .	120
4.4.2	Hepatitis B . . . . .	126
4.4.3	HIV2 . . . . .	128
<b>5</b>	<b>Discussion &amp; Future Work</b>	<b>132</b>
5.1	Annotation of Viruses With Non-Conserved Gene Structure . . . . .	132
5.1.1	Overview . . . . .	132
5.1.2	Introns . . . . .	136
5.1.3	Ribosomal Slippage . . . . .	139
5.1.4	Multiple Sequences . . . . .	140
5.2	Annotation of Selection Strengths in Viruses . . . . .	142
5.2.1	Overview . . . . .	142
5.2.2	RNA secondary structure . . . . .	144

5.2.3	Recombination . . . . .	145
5.2.4	Simultaneous Inference of Alignment . . . . .	145
5.3	Investigating Selection: A Statistical Alignment Approach . .	146
5.3.1	Overview . . . . .	146
5.3.2	Varying Transition and Transversion Rates . . . . .	149
5.3.3	Multiple Alignment . . . . .	149
5.3.4	Organic Choosing of Breakpoints . . . . .	150
5.3.5	Summary . . . . .	151
<b>A</b>	<b>Algorithms</b>	<b>152</b>
A.1	The Viterbi Algorithm . . . . .	152
A.2	The Forward Algorithm . . . . .	153
A.3	The Backward Algorithm . . . . .	154
A.4	The Baum Welch Algorithm . . . . .	154
A.5	The Newton Raphson Iteration . . . . .	155

# Acknowledgements

I would like to thank Professor Jotun Hein for his supervision and the BBSRC for granting me the opportunity of doing this work. I would also like to thank Dr. Thomas Mailund and Dr. Stephen McCauley for their continual support, help and advice throughout this thesis and Miss Naila Mimouni for always bringing a smile to my face. Finally my thanks go out to my parents for their unshakable faith in their offspring and my dog Humphrey, for his critical barks and general wooliness.

# Chapter 1

## Introduction

### 1.1 Abstract

In the past few years we have witnessed an explosion in the viral genomic data available. GenBank alone holds over 80,000 close to complete viral genomes, and numbers are rising fast. For example, since the submission of the first SARS genome in May 2003, over 140 more have been published. With this genomic data at hand we hope to finally be able to tackle our understanding of viruses. Mechanisms of selection and evolution on viruses are still strongly debated, and a methodology which is trimmed towards answering these questions is required. A step towards this is our attempt to develop methods which can deal with the vast amount of viral data, as well as the complexity of viral genomes and their high divergence and subsequent unreliability of alignment.

Several papers have been dedicated to the study of genome annotation and selection on viral genomes, in particular focusing attention on the evo-

lutionary behaviour of overlapping reading frames. This is a feature common to viruses, where due to the three periodicity of the genetic code, up to three genes may be encoded simultaneously in one direction. The constraints placed on a nucleotide involved in such a multiple coding region will naturally have an effect on its mutational pattern, and as a result the concept of selection is complicated further. Additionally, due to their fast evolution time, we observe changes in gene structure between viruses of the same family. Finally, as a result of this high divergence, alignments between two genomes will tend to be unreliable, thus complicating the issue of comparative analysis further.

The focus of this thesis is therefore threefold, because the problems when analysing viral genomes are manifold. We will attempt to fill some of the gaps in the methodology available, creating methods suitable for dealing with annotation both of gene structure and selection strength in viruses. Each chapter of our work relates to one of our publications. We introduce in turn each method, its academic context and its results. We subsequently in chapter 5 discuss for each method its achievements, its shortcomings and future possible extensions and improvements to it.

We first introduce an *ab initio* pairwise comparative annotation method, which not only accounts for the presence of overlapping reading frames in genomes, but also for differences in gene structure between the two compared sequences. Secondly, we develop a hidden Markov model for the annotation of selection strengths across a viral genome accommodating for inter- as well as intragenic differences in selection. Thirdly, we investigate the effect of using a fixed alignment on the inference of selection by incorporating statistical

alignment into our selection analysis.

All three methods presented here improve on their respective equivalents in the field, and we investigate the nature of selection in overlapping regions in several studies on different viral genomes. We begin with a brief introduction to both the biological background and the research up to date done in the field.

## 1.2 Literature Review

### 1.2.1 Biological Background

When looking at the genetic code, there is an underlying three periodicity, where a normal gene is encoded for by a series of nucleotide triplets, each one coding for a particular amino acid. A gene is flanked by a start codon ATG and one of the three stop codons TAG, TGA or TAA. Due to this 3-modularity, we may label through the gene and mark each nucleotide as either coding for the first, second or third position of a codon. We may imagine, however, shifting our labelling one to the right, such that each second position becomes a first, each third a second and each first a third. We would receive a new string of codons, and therefore, if they were flanked in the correct mode by a start and a stop codon, potentially a novel gene. In this case, the region of the sequence coding for two genes simultaneously in different reading frames would be referred to as an *overlapping coding region*.

We can imagine two types of overlap: Firstly, where one gene is completely encased by the other, in which case we refer to it as a *nested* overlapping

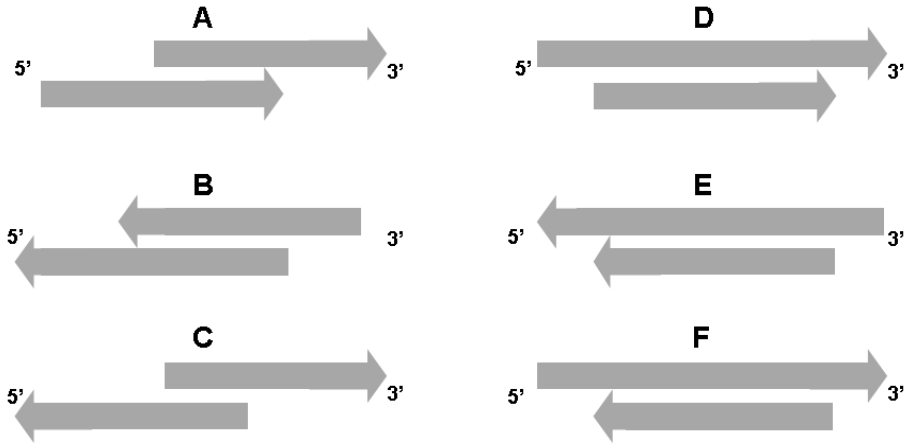


Figure 1.1: The different types of overlap. **A**: unidirectional forward terminal, **B**: unidirectional backward terminal, **C**: bidirectional terminal, **D**: unidirectional forward nested, **E**: unidirectional backward nested, **F**: bidirectional nested.

gene. Secondly, where the 5' and 3' terminal ends of two genes overlap, in which case we refer to them as *terminal* overlapping genes. The majority of genes in organisms get transcribed from the 5' to the 3' end. As is the case in positive sense RNA viruses, it is however possible to code in the other direction, and there are even genomes which are called 'ambisense', such as those of members of the *Bunyaviridae* family. Here part of the genome is coded for in the positive, and part in the negative sense. We may therefore split both the nested and the terminal overlapping genes each into yet again three subdivisions, relating to *unidirectional forward*, *unidirectional backward* and *bidirectional*, where one gene is transcribed in the forward direction and one in the backward direction (see figure 1.1).

The first overlapping gene to have been reported was in bacteriophage  $\phi$ X174 and G4 by Barrell *et al.* [1976] and Fiddes & Godson [1979] respectively, consisting of three entirely nested genes. Following this, many articles

have been published reporting the discovery of overlapping genes in various viruses [Beck *et al.*, 1991, Ding *et al.*, 1994, Giorgi *et al.*, 1983, Mayo *et al.*, 1989, Morch *et al.*, 1988, Pavesi, 2000, Samuel, 1989, Spiropoulou & Nichol, 1993, Walewski *et al.*, 2001].

A number of overlapping genes have also been discovered in higher eukaryotes, such as in *Drosophila* [Misener & Walker, 2000, Misra *et al.*, 2002, Spencer *et al.*, 1986], yeast [Malavasic & Elder, 1990, Peterson & Myers, 1993], mouse [Batshake, 1996, Kasper *et al.*, 2002, Liu *et al.*, 1999, Tvdrick *et al.*, 1999, Williams *et al.*, 1986], and human [Bristow, 1993, Cooper *et al.*, 1998, Duhig *et al.*, 1998, Edgar, 2003, Kennerson *et al.*, 1997, Kiyosawa & Abe, 2002, Laabi *et al.*, 1994, Morel *et al.*, 1989, Nicolaides *et al.*, 1995, Ohinata, 2002, Petruhkin *et al.*, 1998, Zhou & Blumberg, 2003]. However, although overlapping reading frames have been found in a variety of organisms, especially RNA viruses appear to have a tendency towards them [Cann, 1997]. We will therefore focus our research in this body of work on them and now briefly discuss RNA viruses in more detail.

Viruses are sub microscopic parasites that infect cells in biological organisms. They are dependent on the host, in so far as they are incapable of self replication. Normally, an organism encodes for transcriptional machinery in its own genome – RNA viruses in particular however save space by assimilating the transcriptional machinery in the cell of their host for this purpose. RNA viruses have a very small genome which can be either single or double stranded. The genome is packaged in a protein coat, by some viruses further enclosed in a lipid envelope. Additionally, the subgroup of retroviruses carries with it a virus-encoded reverse transcriptase enzyme which enables it to

integrate itself into the host genome to hijack its machinery. A retrovirus is encoded for in the negative direction. It therefore first needs to be reverse transcribed from RNA into DNA with this enzyme, before it can enter the hosts genome using another enzyme called integrase.

RNA viruses have two features which make them stand out in particular: their relatively high mutation rate and their small genome size. Viral polymerases lack a proofreading 3' and 5' exonuclease domain [Steinhauer *et al.*, 1992] and there is additionally no mismatch repair. As a result of this mutation rates are around  $10^{-4}$  per base pair at each replication step [Crotty *et al.*, 2001, Drake & Holland, 1999, Mansky, 2000] which is higher to an order of several magnitudes compared to DNA-based life forms [Drake & Holland, 1999]. Also, the average length of an RNA virus is only 9000 nucleotides, with the smallest ones, such as the *Hepatitis Delta Virus* being of length  $\sim 1700$  nucleotides, ranging up to the longest one such as *Coronaviridae* at  $\sim 32$  kilobase pairs.

Holmes [2003] suggested most recently that the two features of restricted size and high mutation rate were in fact related, due to the high mutation rate actually limiting the genome size of an RNA virus. He argues according to the inverse relationship between the size of any replicating molecule and its mutation rate first laid out by Eigen [1971]. From the latter's argument it would follow, that a virus with a length of 1 Million nucleotides (the approximate length of the largest DNA virus) would suffer from lethal mutations, if it were to have a mutation rate similar to that observed in RNA viruses. This would therefore force RNA viruses to be of shorter length below some sort of error threshold, determined by a function of mutation rate and genome size

as suggested by Nowak [1992]. Jenkins *et al.* [2002] found an inverse proportional relationship between substitution rates and genome size, which would support the idea of shorter viruses being able to have a higher mutation rate for survival than longer ones.

An elegant way to overcome a decrease in informational content of a smaller genome is to encode genes in overlapping reading frames. We could thus expect a larger proportion of overlapping regions in shorter genomes, and indeed Belshaw *et al.* [2007] showed this to be the case with 56% of RNA viruses having some sort of overlap and a significant bias towards these occurring in shorter genomes. One might also assume that overlapping genes were involved in epistasis, that is to say there being an interaction between the genes, such as one suppressing or enhancing the other. A study across 14 viral families by Burch *et al.* [2003] however showed no significant evidence for this phenomenon.

In any genomic sequence mutations occur over time, due to transcriptional error. Some of these mutations will change the amino acid encoded, and subsequently the protein being translated from the relevant gene, and we describe these as *nonsynonymous* substitutions. Other mutations however, may not actually change the translated protein and in this case we refer to them as *synonymous* substitutions. Suppose we are given a reference genome together with an alignment of a set of descendant sequences. When counting the synonymous and nonsynonymous substitutions over time at a particular site, we may create what is known as the  $K_a/K_s$  ratio, which is equal to the number of nonsynonymous substitutions divided by the number of synonymous substitutions observed at that site. If the  $K_a/K_s$  ratio is  $< 1$ , this

means that on average the site prefers a mutation to be synonymous, and thus is inclined not to change the protein it is involved in coding for. We refer to this as the site being under *negative selection*. If however the  $K_a/K_s$  ratio is  $> 1$  this may be seen as an encouragement for a protein-changing substitution, and we refer to this as the site being under *positive selection*. If indeed the  $K_a/K_s$  ratio = 1, then we speak of *neutral evolution*. In an overlapping region however each site codes for two genes simultaneously, so that a synonymous substitution in one reading frame may well be nonsynonymous in the other.

In the case of selectional pressure on an RNA virus generally being negative [Elena *et al.*, 2006, Sanjuan *et al.*, 2004], the effect the creation of an overlapping region has on the fitness of the organism is subsequently twofold: on the one hand it will increase the deleterious effect of each mutation on the organism, since in an overlapping region each mutation might affect two genes. On the other hand, the number of mutations occurring per replication will be reduced, since fewer sites are needed to encode the same genes. A number of papers have focused their attention on the understanding of the composition of nucleotide regions in and the evolution of overlapping reading frames.

### **1.2.2 Sequence Analysis of Overlapping Coding Regions**

Using information theory indices, Pavesi *et al.* [1997] exhibited in a study across several viruses, certain patterns particular to overlapping coding re-

gions. Generally the nucleotide composition tended to be more uniform and the dinucleotide composition more constrained. In particular the usage of highly degenerate codons such as arginine, leucine and serine was noted to be more extreme as compared to non-overlapping regions. To measure this, they introduced the concept of the *RSCU* value:

$$RSCU = \frac{N_{codon}}{N_{aminoacid}} \cdot D \quad (1.1)$$

It was calculated by dividing the number of times a codon was used, by the number of times the amino acid it encodes was coded for in total, multiplied by the degeneracy  $D$  of that amino acid. If all codons were used with equal frequency an *RSCU* value of 1 would be expected. Comparing codon usage in overlapping regions to non-overlapping ones the authors found certain overlapping genes exhibiting a significantly different choice of synonymous codons from that occurring in the corresponding non-overlapping genes. Most interestingly, the authors performed a statistical analysis on the difference of amino acid usage in overlapping versus non-overlapping regions, discovering a strong bias towards the highly degenerate amino acids, such as arginine, leucine and serine in the former. They speculated as to whether the resulting coding flexibility in overlapping reading frames may be viewed as a valuable tool for the evolution of an organism, since the genes would have more liberty to acquire new and maybe specialized functions throughout their life-cycle.

Kozlov [2000,?] investigated the variability of the genetic code within overlapping regions. When considering a four nucleotide sequence, where the inner two nucleotides were double coding, he constructed a table of fre-

quency of overlaps between different amino acids. For example, the 4-mer ACCT would be coding simultaneously for Threonine (ACC) and Proline (CCT). Since, however, Proline is a fourfold degenerate codon ACCA, ACCG and ACCC would be 4-mers which also code for an overlap of Threonine and Proline. He thus constructed a table of relative frequencies between overlapping 4-mers, where clearly the more degenerate a codon is, the more often it featured. He also calculated the expected counts of each overlap frequency based on the degeneracy of the amino acids assuming independence between reading frames. That is to say we would expect to see an overlap between Proline and a particular amino acid  $x$  proportional to the degeneracy of  $x$  if independence between reading frames were to hold, which it obviously does not. Comparing the expected to the observed counts Kozlov noted a preponderance towards the overlap of basic versus acidic residues, such as in the case of Arginine vs. Aspartic and Glutamic Acid. We must remember though, that this is not a feature of overlapping coding regions, but far more a feature of the genetic code. Any sequence of nucleotides would exhibit these features when viewed as a double coding sequence.

Belshaw *et al.* [2007] recently investigated the possible reasons for the creation of overlapping genes and proposed a model for the evolution of gene overlap. In the case of the loss or gain of a stop or start codon respectively, this could result in a terminal overlap, as defined above, and was suggested to be the result of the pressure of genomic compression. A nested overlap was modelled to be the result of a frameshift in an already existing coding region, and the subsequent transcription of an overlapping, yet previously unused, open reading frame. Here Belshaw *et al.* [2007] noted though, that nested

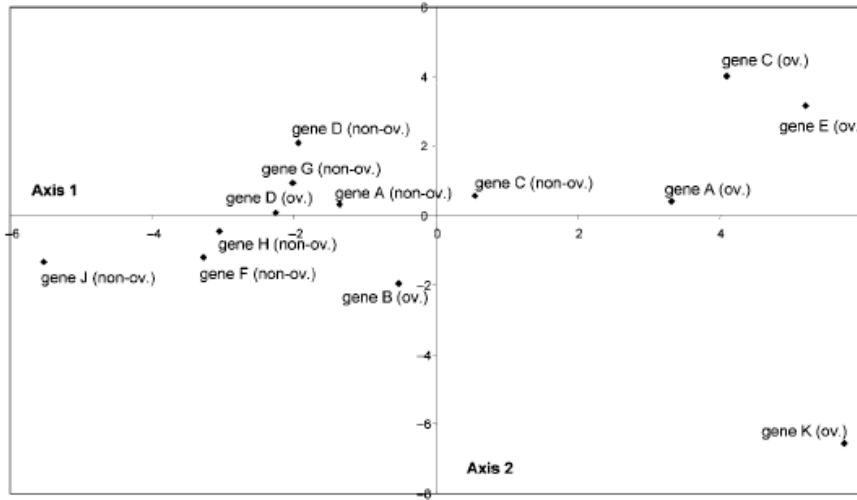
genes are predominantly created by a +1 frameshift. This is believed to be a result of codon usage, common to a wide range of organisms, since coding sequences tend to exhibit a large number of RNY triplet repeats [Shepherd, 1981]. Subsequently, in a +1 and -1 frameshift this would result in a preponderance of NYR and YRN triplets respectively, resulting in us finding a higher proportion of start codons in the +1 and stop codons in the -1 direction. From this it would follow that -1 frameshifts would on average result in smaller open reading frames than +1 frameshifts, and thus potentially be below a certain threshold of being biologically viable. Another supporting aspect to this argument, is that Seligmann & Pollock [2004] even suggest codon usage has evolved in precisely this way to increase the frequency of stop codons in unused open reading frames, in order to minimize damage by accidental frameshifting errors. Belshaw *et al.* [2007] recreated this behaviour in their simulation studies, resulting in a similar distribution of overlaps and frameshifts as empirical results suggested.

### **1.2.3 Investigating Selection Acting on Overlapping Regions**

One of the first articles investigating the nature of selection in overlapping reading frames was by Mizokami *et al.* [1997], who studied the evolution of the overlapping region between the nested *S* and the encompassing *P* gene in Hepatitis B. The authors considered a gene-by-gene alignment of 27 strains. The numbers of synonymous and non-synonymous substitutions for the double coding region were estimated for each gene by using the Nei & Gojobori

[1986] method, and the authors noted that many synonymous substitutions occurred in this region for the  $P$  gene, most of which were non-synonymous for the  $S$  gene. The synonymous substitution rates for the non-overlapping parts of the  $P$  and  $C$  gene were calculated, and turned out to be about 5 times higher than those for the fully nested  $S$  gene. The non-synonymous substitutions however did not show such a dramatic variability. The authors thus concluded the now well-known fact that synonymous substitutions were more constrained in overlapping regions, to a much lesser extent than non-synonymous ones.

Pavesi [2006] considered 30 complete sequences of coliphages, where each sequence was separated into 13 coding segments. Here a new segment began, whenever there was a change in gene structure. In coliphage there are four non-overlapping genes  $J$ ,  $F$ ,  $G$  and  $H$  and three partially overlapping ones  $A$ ,  $C$  and  $D$  in which genes  $B$ ,  $E$  and  $K$  are fully nested. As described in Pavesi *et al.* [1997], he here calculated the  $RSCU$  value for each of the 59 degenerate codons over the 13 different coding segments, and subjected this  $13 \times 59$  matrix to a principal component analysis [Morrison, 1976]. Subsequently the information was condensed down to the use of two axes, as shown in figure 1.2, since these accounted for the vast amount of information. There were two main patterns of codon usage which stood out in particular, corresponding to the overlapping and non-overlapping regions. Pavesi assumed the hypothesis of the synonyms in non-overlapping genes reflecting the ancestral pattern of codon usage. Based on this, he suggested the *de novo* creation of genes  $E$  and  $K$ , since they were furthest apart from the non-overlapping section. This hypothesis was further supported by both the non-overlapping and overlap-



**Fig. 1.** PCA map of the pattern of codon usage in coliphages. Axis 1 accounts for 30 % and axis 2 for 16 % of the total variation in the data matrix. The subsequent axes, accounting for a progressively smaller amount of the residual variation, did not provide more relevant information on the pattern of codon usage (data not shown). Abbreviations indicate overlap (ov.) and non-overlap (non-ov.).

Figure 1.2: The figure shown in the paper by Pavesi [2006]. We can see the clustering to the left of all non-overlapping genes together with both the overlapping and the non-overlapping regions of gene *D*. To the right we see the entirely nested genes *E* and *K*.

ping sections of gene *D* being close together. Gene *B* showed a codon usage similar to the non-overlapping genes, and Pavesi thus hypothesized that gene *A* was originally non-overlapping and developed its current length by using a new stop codon beyond gene *B*. Subsequently alignments of the sequences of the overlapping *A* and *D* genes were made using CLUSTALW [Thompson *et al.*, 1994] and synonymous and non-synonymous substitutions in the overlapping and non-overlapping regions were estimated using Nei & Gojori [1986]. The results revealed selection in the overlapping regions to be negative in one gene and positive in the other.

Rogozin *et al.* [2002] also investigated selection on genes coding in overlapping reading frames. Since their formation necessarily involved creating

a coding region from non-coding DNA, they hoped to thus learn more about *de novo* gene creation. A set of 71 trusted pairs of overlapping bidirectional genes from prokaryotes was analysed, where the 3' end overlapped by more than 15 nucleotides. As the dominating mode of evolution in such multiple coding regions they investigated the three principal scenarios:

1. The new protein coding region, particularly the C-terminal, was under no functional constraint, and thus under neutral selection.
2. The new region was under positive selection.
3. The modes of evolution in the two terminal regions could differ, i.e. the old coding sequence could be undergoing negative and the new one positive selection.

Rogozin *et al.* [2002] looked at an overlap of two genes conserved over six members of the *Chlamydiaceae* family. They conducted a test for purifying selection to assess the importance of functionality of the overlapping regions. In single coding sequences one can take a low non-synonymous to synonymous ratio as an indication for negative selection; however in multiple reading frames this distinction is frame-dependent. They thus concentrated on fourfold degenerate third positions of overlaps in phase (123:132). By doing so the second codon position was opposite the third fully degenerate one and thus under no mutational constraint in the other reading frame. They found a significantly lower substitution rate in overlapping regions than in non-overlapping, indicating purifying selection being at work in the former. Secondly, they noted that more than half the overlaps found were in phase

(123:132), which was a significant deviation from the expected distribution. Since this coding phase permits the most amino acid replacements in one overlapping region without affecting the other one, the authors used this as an indication for the favourability of positive selection, at least for the early stages of the evolution of the new protein sequences.

Guyader & Ducray [2002] investigated the Potato leafroll virus in detail and, amongst other things, selection on the overlapping regions therein. Twelve strands from different countries were taken and aligned via CLUSTALW [Thompson *et al.*, 1994]. Subsequently several different codon-based Markov models of substitution from the PAML package [Yang *et al.*, 2003] were used to analyse the  $K_a/K_s$  ratio on the different regions. The models tested included

Model	Description
M1	Two categories: $\omega = 0$ and $\omega = 1$
M2	Three categories: $\omega = 0$ , $\omega = 1$ and $\omega > 1$

For a certain overlap, the authors found a significantly better fit of model M2 to one reading frame ( $P < 0.05$ ), whereas M1 could not be rejected for the other one. From this they concluded that differential selection was occurring on this overlap, where one gene was under significantly stronger negative selection than the other.

An analysis of 7 SIV sequences was performed by Hughes *et al.* [2001]. An alignment was obtained via CLUSTALW at the amino acid level and the number of synonymous and non-synonymous substitutions were estimated by

the method introduced by Nei & Gojobori [1986]. The authors noted that in the region of overlap between the *tat* and *vpr* gene, non-synonymous substitutions in the *tat* gene occurred in such a way as to cause mostly synonymous substitutions in the *vpr* gene. Additionally the *vpr* gene demonstrated signs of stronger negative selection than the *tat* gene, based on a lower estimate of non-synonymous substitutions. The authors also found, that contrary to the usual scenario where the number of synonymous substitutions across the genome is relatively constant, in genomes with overlapping coding regions this is far from the case. Here the number of synonymous substitutions within a gene is closely related to the extent to which that gene overlaps another. However, even in spite of this reduction in rate the authors still observed strong evidence for the detection of positive selection on a certain epitope within the *tat* gene.

When analysing a CLUSTALW alignment of 22 Human Papillomavirus sequences Narechania *et al.* [2005] estimated a high overall  $K_a/K_s$  ratio [Nei & Gojobori, 1986] in the E2 gene, which contains the nested E4 gene. The E2 gene appeared to favour non-synonymous changes, whereas the nested E4 gene tended towards synonymous changes and thus greater conservation.

Almost simultaneously Hughes & Hughes [2005] published a study similar to the one described above by Narechania *et al.* [2005] on an alignment of HPV viruses. They confirmed the observation of positive selection on the overlapping region of the E2 gene coexisting with negative selection on the E4 gene.

### 1.2.4 Modelling Selection on Overlapping Regions

Hein & Støvlbæk [1995] developed a model describing the evolution over time of sequences coding in overlapping reading frames. This allowed them to analyse combinations of non-coding, singly coding and multiple coding regions of two aligned homologous DNA sequences. The model extended Kimura's two parameter model, adding a selection factor onto replacement substitutions. With the intent of estimating these selection factors and transition/transversion rates, the authors introduced a likelihood function to test a hierarchy of hypotheses of varying strength. The model assumed that transitions and transversions occurred at a rate of  $\alpha$  and  $\beta$  respectively. Assuming an evolutionary distance of  $t/2$  to the most recent common ancestor, the expected number of transitions and transversions per site is  $a = \alpha t$  and  $b = \beta t$ . Additionally the authors assumed that any replacement substitution has a probability  $f$ , the selection factor, of being accepted ( $f < 1$  would imply negative and  $f > 1$  positive selection). Looking at each nucleotide, in each reading frame context, they marked it as one of three types (1:1:1:1, 2:2, or 4) depending on its level of degeneracy. E.g. if they had the sequence CTATC, then following the middle nucleotide through the three potential reading frames we would have CTx = 4, TxT = 1:1:1:1 and xTC = 1:1:1:1. Based on this the authors developed the following extension to Kimura's notation:

- Let  $X(a, b)$  be the probability that a position is identical in both sequences.
- Let  $X_t(a, b, f)$  be the probability that a position of type  $t$  is identical

in both sequences, i.e.

$$X_{1:1:1:1}(a, b, f) = X(a \cdot f, b \cdot f) \quad (1.2)$$

$$X_{2:2}(a, b, f) = X(a, b \cdot f) \quad (1.3)$$

$$X_4(a, b, f) = X(a, b) \quad (1.4)$$

- Define  $Y, Y_t, Z$  and  $Z_t$  similarly for transition and transversion.
- Let  $a, b$  be transition and transversion rates and  $f$  the selection factor as above.

They thus created the likelihood function of a certain alignment given the evolutionary model, where  $x(t)$ ,  $y(t)$  and  $z(t)$  were the occurrences of an identity, transition and transversion respectively at a site of type  $t$ .

$$\text{Likelihood}(\text{alignment}, a, b, f) = \prod_t X_t(a, b, f)^{x(t)} \cdot Y_t(a, b, f)^{y(t)} \cdot Z_t(a, b, f)^{z(t)} \quad (1.5)$$

Thus with a trusted alignment and annotation they suggest using maximum likelihood methods to find good estimates for selection factors and transition and transversion rates. The question then addressed was whether one should have a different selection factor for different genes  $A$  and  $B$ , say  $f_A$  and  $f_B$ . And if so, must one have a new selection factor  $f_{AB}$  for a region encoding both these genes or may one assume they are independent and merely take the product  $f_A \cdot f_B$ ? The authors thus reached various conclusions by means of hypothesis testing, including that independence of selection factors was acceptable.

Pedersen & Jensen [2001] improved on Hein & Støvlbæk [1995]'s work by presenting a model for the substitution process of nucleotides in double coding sequences, and accounting for the change in codon context over time. The authors started off by considering an alignment of two homologous DNA sequences, descended from a common ancestor, which encoded two genes simultaneously, in reading frames  $I$  and  $II$ . They let  $z_i = (z_i^1, z_i^2, z_i^3)$  be the  $i^{th}$  codon in reading frame  $I$ , where  $z_i^k$  was the nucleotide in codon position  $k$  ( $k = 1, 2, 3$ ). They subsequently defined  $\bar{z}_i$  to be a codon that differs from  $z_i$  in one nucleotide position only, and the instantaneous substitution rate from codon  $z_i$  to codon  $\bar{z}_i$  as proportional to  $\pi(\bar{z}_i) = frequency(\bar{z}_i)$ . They defined  $f_I$ ,  $f_{II}$  and  $f_{I/II}$  to be the selective constraints imposed on a mutation that results in a codon change in reading frame  $I$ ,  $II$  and both respectively. They let  $q(z_i, \bar{z}_i | z_{i-1}^2, z_{i-1}^3, z_{i+1}^1)$  be the instantaneous rates of substitution from a sequence that had  $z_i$  as its  $i^{th}$  codon in reading frame  $I$  to a sequence that was identical except for holding  $z_i$  as its  $i^{th}$  codon, at an instant where codon positions 2 and 3 of codon  $i - 1$  and codon position 1 of codon  $i + 1$  are  $z_{i-1}^2$ ,  $z_{i-1}^3$  and  $z_{i+1}^1$  respectively. They subsequently described the model by the following equations:

$$q(z_i, \bar{z}_i) = \begin{cases} 0 & \text{if } z_i \text{ and } \bar{z}_i \text{ differ by more than one nucleotide} \\ q(z_i, \bar{z}_i | z_{i-1}^2, z_{i-1}^3, z_{i+1}^1) & \text{if } z_i \text{ and } \bar{z}_i \text{ differ at one position in codon } i \end{cases}$$

where,

$$\begin{aligned}
q(z_i, \bar{z}_i | z_{i-1}^2, z_{i-1}^3, z_{i+1}^1) &= 0, & \text{STOP,} \\
&= K\pi(\bar{z}_i), & \text{no STOP, ts, syn(I), syn(II)} \\
&= K\pi(\bar{z}_i), & \text{no STOP, tv, syn(I), syn(II)} \\
&= f_I K\pi(\bar{z}_i), & \text{no STOP, ts, non(I), syn(II)} \\
&= f_I K\pi(\bar{z}_i), & \text{no STOP, tv, non(I), syn(II)} \\
&= f_{II} K\pi(\bar{z}_i), & \text{no STOP, ts, syn(I), non(II)} \\
&= f_{II} K\pi(\bar{z}_i), & \text{no STOP, tv, syn(I), non(II)} \\
&= f_{I/II} K\pi(\bar{z}_i), & \text{no STOP, ts, non(I), non(II)} \\
&= f_{I/II} K\pi(\bar{z}_i), & \text{no STOP, tv, non(I), non(II)}
\end{aligned}$$

The authors then described a Markov Chain Monte Carlo simulation technique for obtaining the transition probabilities from one sequence to another. Since the instantaneous rates of substitution under the model depend on the states at neighbouring sites at the time of substitution, they could not obtain the probability of the full transition as a product of marginal transition probabilities. The substitution processes at all sites must instead be considered simultaneously. The Markov chain was initialized with a certain path  $P$ , which is a collection of paths  $P_i$  of the  $i^{\text{th}}$  codon in reading frame  $I$ . Then they essentially searched through the pathspace  $P_i$  for each codon  $i$ , to propose new paths  $P_i^*$ , consistent with the paths  $P_{i-1}$  and  $P_{i+1}$ . The most likely one according to some threshold was then accepted and the full path  $P$  updated with it. Parameters were obtained using maximum likelihood estimation and a simpler multiplicative model, where the selection factor  $f_{I/II}$  was replaced by the product  $f_I \cdot f_{II}$ , was suggested and accepted as a good approximation. By extending context dependency one nucleotide to the right, one may easily extend this model to accommodate for the constraints imposed by three overlapping coding regions, and combining different types of model for the non-, single-, double- and triple coding state allows one to model an entire

genomic region. The authors were justifiably confident about the accuracy of their model; however a main drawback of the entire method is its immense computational time requirement, limiting its practical use considerably.

### 1.2.5 Gene Annotation in General

One aspect we will be dealing with in our research is the problem of gene annotation on multiple coding genomes. We will initially give a brief introduction to the research up to date, which deals with gene finding in general and which has influenced our methods.

Pedersen & Hein [2003] described a probabilistic model of both genome structure and evolution, called an EHHM (Evolutionary Hidden Markov Model) which can be used on any number of single-coding multiply aligned genomes. It consists of an HMM and a set of region-specific evolutionary models based on a phylogenetic tree, all of whose parameters are estimated by the Maximum likelihood method. The model is used for gene annotation in both simulated data and on a set of orthologous human/mouse gene pairs. The authors devised the EHMM, by letting every state  $k$  have an alphabet  $C_k$  over alignment columns, and an emission distribution  $e_k$  specified by a state-specific evolutionary model  $E_k$  and a phylogenetic tree  $T$ . Thus the probability of observing a particular alignment column  $c$  in state  $k$  equals

$$e_k(c) = P(c|E_k, T) \tag{1.6}$$

However, since only the branch lengths and outer nodes of the tree  $T$  are known, the authors had to sum over all possible character states of the inner

nodes. As an evolutionary model the HKY model [Hasegawa *et al.*, 1985] for the single nucleotide states, and the Goldman & Yang [1994] model for the triplet states was used. Parameter estimations were made using the Baum-Welch algorithm [Durbin *et al.*, 1998]. When testing their model on simulated data, the authors reported a rise in performance with both increasing tree length and  $dN/dS$  ratio, as well as with increasing number of genomes compared. For each of four disjoint subsets of the human/mouse genome alignment, a set of model parameters for the EHMM was estimated. Estimates for the evolutionary parameters between the coding and non-coding state showed a significant difference, thus indicating the strong justification of the above approach. Their experiments showed that gene finding can benefit from an EHMM approach when homologous sequences are available. However the simple model they introduced was not able to compete with state of the art gene finders such as GENSCAN. Instead they suggested extending these existing models to an EHMM and thereby using the available evolutionary information to increase performance.

Meyer & Durbin [2002] presented a novel comparative method for *ab initio* prediction of gene structure in eukaryotic genomes using a pair General Hidden Markov Model. Their program DOUBLESCAN both aligns and annotates two eukaryotic genomes with their genome sequences as its only input. The algorithm can model partial genes, multiple genes, single complete genes or no genes, and additionally can align more diverged genes which have been subject to exon-fusion or exon-splitting. The HMM consists of 54 states, split into the following classes, where classes 3-6 each contain three states belonging to match, delete and insert:

1. Begin and End states
2. START/START and STOP/STOP states
3. Exons
4. Splice Sites and Introns with translated regions
5. Splice Sites and Introns with untranslated regions
6. Intergenic states

In each state  $s_i$  a sequence of length  $k_i$  and length  $p_i$  gets emitted for genome 1 and 2 respectively. The probability of a certain state path  $S$  given sequences  $X$  and  $Y$  is therefore given by

$$P(X, Y, S) = e_{s_1}(k_1, p_1) \cdot \prod_{i=1}^{Z-1} a_{s_i, s_{i+1}} e_{s_{i+1}}(k_{i+1}, p_{i+1}) \quad (1.7)$$

Using the Viterbi Algorithm (and a newly introduced algorithm called the “Stepping Stone Algorithm”), the authors thus retrieved the optimal state path through the sequence, and thereby aligned and annotated simultaneously due to the nature of their model. Considering a test set of 80 orthologous mouse and human DNA sequences, they compared the performance of DOUBLESCAN to that of GENSCAN. They found a 10% and 4% higher performance in sensitivity and specificity, respectively, than GENSCAN. Additionally the comparative performance of DOUBLESCAN increased progressively when going from a smaller nucleotide scale to a larger genome wide one.

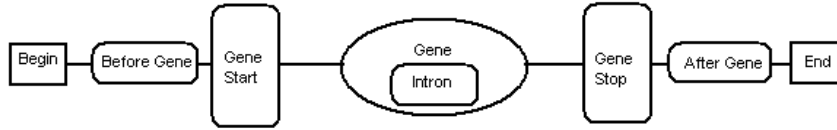


Figure 1.3: Picture of the states in the Hobolth and Jenses model.

Hobolth & Jensen [2005] introduced an HMM that allowed them to analyse comparatively multiple sequences related by a phylogenetic tree. By incorporating methods of structure prediction, statistical multiple alignment and phylogenetic information, this model proved particularly useful for a detailed characterization of homologous DNA sequences sharing a common gene. The basic idea is to have the following hidden Markov chain of states:

The Before, After and Intron Gene states emit Match, Delete or Insert pairs of single nucleotides according to the TKF91 model [Thorne *et al.*, 1991]. The substitution probabilities are determined by the HKY model [Hasegawa *et al.*, 1985] and Goldman and Yang Goldman & Yang [1994] model, in the single nucleotide and codon cases respectively. The Gene Start state emits two aligned ATGs. The Gene state emits Match, Delete or Insert pairs of nucleotide triplets according to the TKF91 model. The Gene

Match A,B	Delete A,B	Insert A,B	Start, Match C, Stop	Delete C	Insert C
#	#	-	###	###	- - -
#	-	#	###	- - -	###

Stop state emits two aligned STOP codons. The parameters of the model were estimated by a modified version of the EM algorithm. Finally, the gene structure prediction was obtained from the Viterbi algorithm with the EM-estimated parameters as an input. Due to the nature of the TKF91 model, this clearly also provided a statistical alignment of the two sequences. Fur-

thermore the authors demonstrated how to extend the above methodology to deal with multiple sequences, by expressing the TKF91 model as a hidden Markov chain along any number of sequences related by a phylogenetic tree (viz. Pedersen & Hein [2003]).

### 1.2.6 Gene Finding in Overlapping Reading Frames

Firth & Brown [2005] developed several statistics to be used on a pairwise alignment, including a new maximum likelihood method, for testing whether a certain region is double coding. For two aligned homologous sequences  $S_1$  and  $S_2$  define the following

*$N_{123}$* : *mutation rate in  $N_1, N_2, N_3$*  For the primary reading frame the number of nucleotide differences between the two sequences in the first, second and third nucleotide position —  $N_1, N_2$  and  $N_3$  — are counted. These are then expressed as a fraction of the total number of  $N_1, N_2$  and  $N_3$  loci and these statistics are labelled as  $f_{N_1}, f_{N_2}$  and  $f_{N_3}$ .

*$NsNn$* : *synonymous and non-synonymous substitutions* For the primary reading frame the number of identical, synonymous and non-synonymous codon substitutions between the two sequences are counted. The number of synonymous and non-synonymous substitutions is expressed as a fraction of the total number of codon pairs and these statistics are labelled  $f_{syn}$  and  $f_{non}$ .

*$MLOGD$* : *maximum likelihood method* The  *$MLOGD$*  (Maximum Likelihood Overlapping Gene Detector) method estimates the relative probabilities

that  $S_2$  descended from  $S_1$  under the single or double coding hypothesis.

A likelihood ratio test is then used to decide between the two.

If one models the evolution of a single sequence as a Markov process, then one may express the probability of  $S_1$  evolving into  $S_2$  by

$$\log P(S_1 \rightarrow S_2; t) = \sum_k^{N_{codons}} \log P(C_1^k \rightarrow C_2^k; t)$$

where  $C_1^k$  and  $C_2^k$  are the  $k^{th}$  codons in sequence  $S_1$  and  $S_2$  respectively, and  $\mathbf{P}(t) = \exp(\mathbf{R}t)$ , where  $\mathbf{R}$  is a  $64 \cdot 64$  matrix of instantaneous codon mutation rates. If one however is dealing with double coding sequences, then one may no longer assume an independence of codons, disallowing one to perform the above factorization. Instead the authors introduced a  $64 \cdot 1$  codon usage matrix  $C$ , and a  $20 \cdot 20$  amino acid substitution matrix  $A$ , additional to  $\mathbf{P}(t) = \exp(\mathbf{Q}t)$  where  $Q$  is a  $4 \cdot 4$  nucleotide substitution matrix. Let the nucleotide pair  $N_1^k, N_2^k$  be coding in the primary reading frame for amino acid  $X_1, X_2$  in  $S_1, S_2$  respectively. For each nucleotide pair they now estimated the probability that  $N_1^k$  mutates into  $N_2^k$  under both the single and the double coding model, and deduced from that the full probability of one sequence having developed into the other under both models. A simple likelihood test

then revealed which model fits the data better. So,

$$b(N_1^k \rightarrow N_2^k; t, \textit{single}) = \mathbf{P}(N_1^k \rightarrow N_2^k; t) \cdot \mathbf{C}(X_2) \cdot \mathbf{A}(X_1 \rightarrow X_2) \quad (1.8)$$

$$b(N_1^k \rightarrow N_2^k; t, \textit{double}) = \mathbf{P}(N_1^k \rightarrow N_2^k; t) \cdot \mathbf{C}(X_2) \cdot \mathbf{A}(X_1 \rightarrow X_2) \\ \cdot \mathbf{C}(X_2') \cdot \mathbf{A}(X_1' \rightarrow X_2') \quad (1.9)$$

$$\mathbf{P}(N_1^k \rightarrow N_2^k; t, x) = \frac{b(N_1^k \rightarrow N_2^k; t, x)}{\sum_{i=A,C,T,G} b(N_1^k \rightarrow i; t, x)} \quad (1.10)$$

$$\log \mathbf{P}(S_1 \rightarrow S_2; t, x) = \sum_N \log \mathbf{P}(N_1^k \rightarrow N_2^k; t, x) \quad (1.11)$$

$$(1.12)$$

Using the above three statistics on known overlapping and non-overlapping coding regions the *MLOGD* turned out to be much more reliable than the prior two methods, giving accurate results of around 80%. One main drawback of the method is that it can only be used to classify the coding nature of a certain region in question, but is unable to annotate entire genomes.

Following this, Firth & Brown [2006] developed the *MLOGD* method into a software program, presented a database of results for 640 virus sequence alignments and incorporated this into an online interface. Their method now is extended to a multiple sequence method, by constructing a phylogenetic tree with PHYLIP [Felsenstein, 1989] and using it to create a list of sequence pairs by tracing around the perimeter of the tree, so that each sequence gets used exactly twice. Subsequently the *MLOGD* is calculated for each set of pairs, summed over all the pairs and divided by 2 to give the *MLOGD* of the multiple sequence alignment. As well as testing whether an overlapping open reading frame in question is coding or not, one can also perform *ab initio*

testing by investigating all open reading frames above a specifiable length. One may also apply a six-frame sliding window test to discover functional regions, which due to ribosomal slippage or frameshifting may not begin with a start codon.

In their very recent paper, McCauley *et al.* [2007] introduced the novel idea of an HMM framework to explicitly annotate a single viral genome coding in multiple reading frames. They extended the idea of a sequence being generally coding or non-coding to it being so in each reading frame, resulting in an 8 state conditional HMM. Making the state-dependent emission probabilities additionally conditional on the prior two emitted nucleotides, they managed to incorporate the bias of multiple coding regions towards highly degenerate amino acids into their HMM framework. Let us consider  $C(i|N)$ ,  $C(i|S)$ ,  $C(i, j|D)$ ,  $C(i, j, k|T)$  to represent the  $i^{th}$ ,  $j^{th}$  and  $k^{th}$  coding position loci ( $i, j, k = 1, 2, 3$ ) in a non-, single-, double- and triple coding region respectively. So a nucleotide in a double coding region, coding at position 1 in one reading frame, and position 3 in another would be denoted by  $C(1, 3|D)$ . Then the state-dependent emission probability  $e_i$  of emitting the  $i^{th}$  nucleotide, is drawn from one of  $x$  ( $x = 1, 4, 16$ ) multinomial distributions where  $x$  is dependent on the state. We may summarize this as seen in Table 1.1. Transition probabilities are conditional on occurrence of a START codon and happen at rate  $\alpha$ . The multinomial distributions are subsequently found using an Expectation Maximization procedure and the most likely state annotation is found with the Viterbi Algorithm. Following this the authors extended the model to a Phylogenetic EHMM, as introduced for gene finding by Pedersen & Hein [2003]. Using the evolutionary infor-

State	Conditional On	x	Number of Parameters
$C(0)$	none	1	1*4
$C(1 S)$	none	1	1*4
$C(2 S)$	$e_{i-1}$	4	4*4
$C(3 S)$	$e_{i-1}, e_{i-2}$	16	16*4
$C(1, 2 D)$	$e_{i-1}$	4	4*4
$C(1, 3 D)$	$e_{i-1}, e_{i-2}$	16	16*4
$C(2, 3 D)$	$e_{i-1}, e_{i-2}$	16	16*4
$C(1, 2, 3 T)$	$e_{i-1}, e_{i-2}$	16	16*4

Table 1.1: The different state assignments in the McCauley & Hein Model

mation that a multiple alignment with homologous sequences provides, they improved significantly on their prior annotations. Finally McCauley & Hein compared both methods to GenMark and found highly encouraging results, confirming that their HMM approach was indeed a fully justified one — especially for more complex genomes — since it allowed an interaction between the different reading frames.

Building on his earlier work, Pavese [2000] introduced a novel feature particular to overlapping reading frames. By analysis of a sample of 21 viral genomic sequences he discovered the occurrence of a cluster of basic amino acid residues encoded in one frame, overlapping with a cluster of acidic residues in another reading frame. Using this criterion the author predicted a novel overlapping gene in the Hepatitis G Virus. Additionally the proposed region demonstrated a very low rate of synonymous substitutions supporting the presence of a multiple coding region. Using a  $k$ -tuple statistical analysis a  $\chi^2$  contingency table test was carried out to assess whether significant differences existed between the compositions of overlapping and non-overlapping coding sequences. The test was carried out at the individual amino acid

( $k = 1$ ), dipeptide ( $k = 2$ ) and tripeptide ( $k = 3$ ) level, and the bias towards synonymous codons was tested by estimating the “effective number of codons” for a given amino acid residue. Pavesi thus found that

- All viruses show a significantly high level of Arginine and Leucine in the overlapping coding regions
- The use of Leucine synonyms in overlapping reading frames encoding polyleucine genes is similar to that in non-overlapping ones
- The use of Arginine synonyms in overlapping reading frames is however more biased than that occurring in non-overlapping frames
- The content of AGA and CGA codons in overlapping frames expressing Arginine clusters is always higher than that of the respective non-overlapping set
- There will therefore be a motif consisting of a high number of acidic residues in a reading frame overlapping with a cluster of AGA and CGA codons in another reading frame

Using the above motif the author introduced an algorithm to scan a database for multiple coding regions, and discovered a potential new overlapping gene in Hepatitis G Virus, additionally supported by other factors. However, a main drawback of the described method is it resulting in a relatively high number of false positives, probably due to the motif being quite short.

In a similar vein, Walewski *et al.* [2001] detected an unusually conserved region on the Hepatitis C virus antigen. The authors studied eight highly divergent sequences and noticed fourfold degenerate unusually conserved sites

at the third codon positions in a known gene. All eight sequences contained the same glycine codon GGA, whereas GGG, GGT and GGC also code for the same amino acid. The probability of such conservation under the assumption of no cost to a mutation in the third position would be one in  $4^7 = 16,384$ . A logical explanation was therefore for evolutionary constraints to be imposed by a gene coding in a different reading frame. Several other clusters of conserved codons were found in neighbouring regions. Around these clusters the authors searched for unused alternative open reading frames, defined as a stretch of nucleotides of length  $> 50$  without an in-frame stop codon. A long open reading frame of length  $> 124$  was found and clinical tests proved this indeed to be a functional region.

### 1.2.7 Decreasing Uncertainty when Using Alignments

Metzler *et al.* [2001] developed a method for the estimation of mutation rates without a bias towards a fixed alignment. By performing the joint sampling of alignments and mutation rates, they obtain a more realistic idea of the uncertainty underlying their parameter estimates. They use the TKF91 model [Felsenstein, 1989] for two sequences  $S_1, S_2$  with parameters  $\Theta$  given by the substitution rate  $s$ , insertion rate  $\lambda$  and deletion rate  $\mu$ . If  $\Theta$  were fixed they could sample an alignment applying classical HMM backward sampling algorithms such as described in Durbin *et al.* [1998]. Similarly if an alignment were given they would use a Metropolis-Hastings approach for sampling  $\Theta$  (see Gamerman [1997]). The authors combine these two ideas by using the idea of Gibbs Sampling [Gamerman, 1997] and obtain a method for sam-

pling mutation parameters and alignment simultaneously. Thus they present a novel method to assess joint variability of alignment and parameter estimation, and decrease uncertainty due to the use of an 'optimal' alignment.

A more practical application is found by Lunter *et al.* [2004]. The author introduces a method for indel estimation which sums over all alignments using marginalized posterior decoding. He shows in a simulation study that estimates for indel rates are within 2% of the simulated parameter value. Subsequently a study of the human and mouse genomes reveals that indel rates appear to be up to twice as high as suggested by the use of a fixed alignment.

## Chapter 2

# Annotation of Viruses with Non-Conserved Gene Structure

*In this chapter we discuss the work published in de Groot et al. [2007]. Our research focuses on the yet unaddressed question of comparative genome annotation for viruses, where the gene structure differs between the two compared sequences, whilst still accounting for overlapping reading frames — both from a structural and an evolutionary perspective. The model was fully developed and programmed up by myself, as well as all experiments and simulations run by me. Dr. Thomas Mailund helped in presenting the work in a publishable form.*

## **2.1 Abstract**

### **2.1.1 Motivation**

Detecting genes in viral genomes is a complex task in particular due to the tendency of coding in overlapping reading frames. Conventional 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 a hidden Markov model 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.

### **2.1.2 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 about 84–89% and specificity of about 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.

## 2.2 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 where one is fully nested 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 have often undergone much evolution and gene structure might have changed significantly over time. For example, HIV1 and HIV2 have nine genes each, however only 8 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 a hidden Markov model 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.

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 & Brown, 2005].

Building on the evolutionary model introduced by Hein & Støvlbæk [1995], we demonstrate how to extend it to a hidden Markov model 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 8 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 dominate our results.

## 2.3 Methods

### 2.3.1 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  $\Sigma$ , 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  $\Sigma$ . 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  $\pi$  is given by

$$P(x, \pi | A, B, e) = B(\pi_1) e_{\pi_1}^{x_1} \prod_{k=2}^K e_{\pi_k}^{x_k} a_{\pi_{k-1}, \pi_k} \quad (2.1)$$

There are three unidirectional global reading frames, fixed before annotation of the sequence, which will henceforth be 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 figure 2.1). 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  $\Sigma$  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 a hidden Markov model 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 (see Figure 2.2). Additionally, we set the silent start and end states to be non-coding in all three reading frames, thus ensuring our annotating only ‘entire’ genes as coding.

### 2.3.2 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

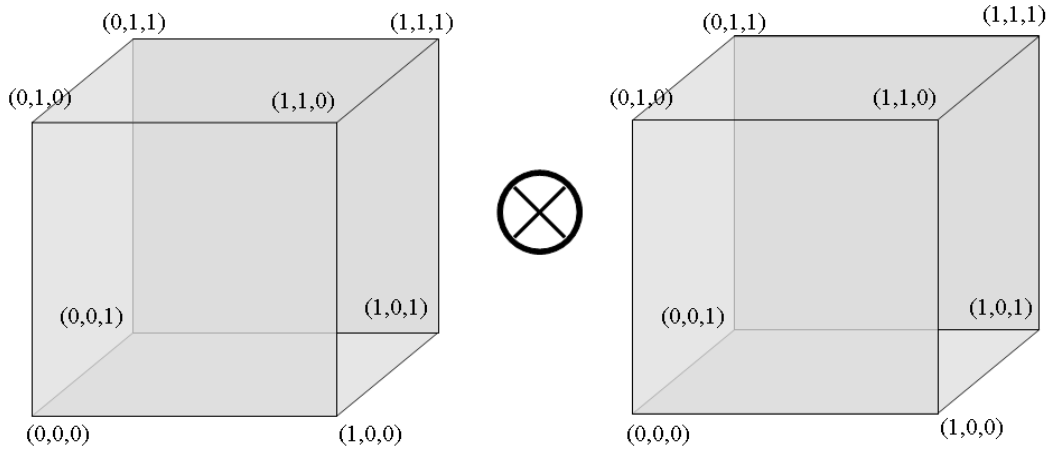


Figure 2.1: The hypercube representing the 64 states the two sequences can jointly be in. The vertices of the left cube represent the eight states the first sequence may be in. Here  $(0,0,0)$  is non-coding, whereas  $(1,1,1)$  is triple coding. Similarly the right cube represents the states of the 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  $8 \times 8$  combinations of the ‘cross product’ of the two cubes.

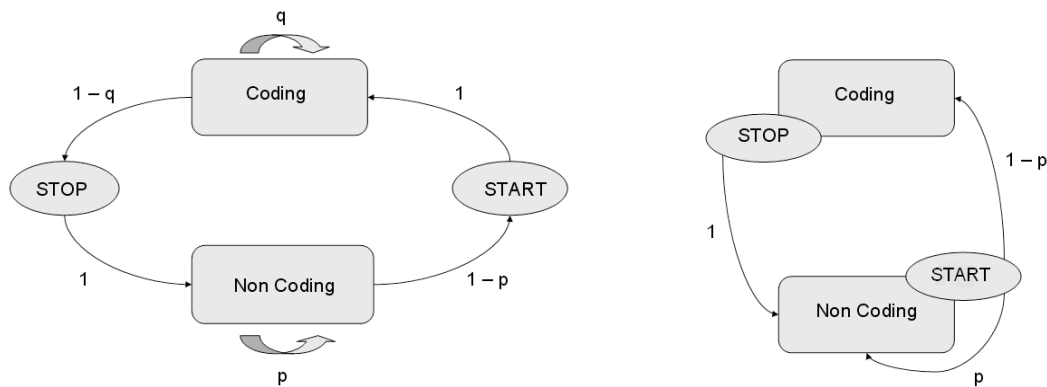


Figure 2.2: To the left the standard HMM, which switches into coding via the START state. When in the non-coding state it emits a start codon with probability  $1 - p$ ; when in the coding state it emits a stop codon with probability  $1 - q$ . To the right the conditional version which switches from non-coding into coding with probability  $p$  **conditional** on observing a start codon.

- (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. Representing these states by the eight vertices of a cube additionally visualizes the restriction of only being allowed to move from one state to another along the edge of a cube, when walking through the sequence alignment. We thus constrain some of the state transition probabilities to being zero, e.g. we may not transition from non-coding to double-coding in one step, since at every nucleotide position only one new reading frame is in question. Since  $|S| = 64$  we could have  $64^2$  transition probabilities, many of which however would be zero due to the constraints mentioned above. However, 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 (figure 2.3), conditional on finding a start codon in the respective sequence:

- Both sequences are non-coding in GRF  $x$ , we scan an aligned ATG in GRF  $x$  in both and both switch to coding — transition probability  $\alpha$ .

- Both sequences are non-coding in GRF  $x$ , we scan an ATG in one sequence in GRF  $x$  but not in the other and that one switches to coding — transition probability  $\beta$ .
- One of the sequences is already coding in GRF  $x$ , we scan an ATG in the other in GRF  $x$  and it switches into coding as well — transition probability  $\gamma$ .

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.

So without loss of precision let us assign to each state  $s$  in  $S$  one of the 64 vertices of the hypercube (2.1). Let  $y_0$  be the zero vertex  $(0,0,0)$  and  $y_1$ ,  $y_2$ , and  $y_3$  be the three base vectors in three dimensions  $(1,0,0)$ ,  $(0,1,0)$  and  $(0,0,1)$ . We may then describe each state  $s$  as a pair of vectors  $(x_i, x_j)$  where  $x_i$  and  $x_j$  represent sequence 1 and 2 respectively and must each be one of the eight vertices on the unit cube (see again figure 2.1). Let  $m = 1 + l \pmod 3$  and  $\cdot$  represent the vector dot product. Note that if  $x_i \cdot y_m = 0$  then sequence 1 is **not** coding in gRF  $m$  and if  $x_i \cdot y_m = 1$  then sequence 1 **is** coding in gRF  $m$ . Then when looking at the  $l^{th}$  position in the sequence we may write the conditional transition probabilities to the certain states as follows. Remember that since at each locus we may maximally transition into one other state, so the probability of not transitioning, i.e.  $P((x_i, x_j) \rightarrow (x_i, x_j))$ ,

will be  $1 - \alpha$ ,  $1 - \beta$  or  $1 - \gamma$  depending on the context.

- If ATG in both sequences

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i + y_m, x_j + y_m)) = \alpha$

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i + y_m, x_j)) = \gamma$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j + y_m)) = \gamma$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If ATG only in first sequence

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i + y_m, x_j + y_m)) = \beta$

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i + y_m, x_j)) = \gamma$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If ATG only in second sequence

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j + y_m)) = \beta$

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j + y_m)) = \gamma$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If TAG/TGA/TAA in both sequences

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

- If  $x_i \cdot y_m = 0 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j - y_m)) = 1$

- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i - y_m, x_j)) = 1$
- If  $x_i \cdot y_m = 1 \wedge x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i - y_m, x_j - y_m)) = 1$
- If TAG/TGA/TAA only in first sequence
  - If  $x_i \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i - y_m, x_j)) = 1$
  - If  $x_i \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$
- If TAG/TGA/TAA only in second sequence
  - If  $x_j \cdot y_m = 1$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j - y_m)) = 1$
  - If  $x_j \cdot y_m = 0$ ,  $P((x_i, x_j) \rightarrow (x_i, x_j)) = 1$

Keeping in mind that we do not model sequence length or draw sequence composition into account, we will most likely find short regions of high conservation along the genome. It is therefore not unsurprising to note that this will result in the identification of many short false positive reading frames. To counteract this, we condition the probability on a new coding region starting on there not being a stop codon in the respective reading frame within the next  $l$  nucleotides (default value 50). Note that this is not as ‘manual’ as merely removing all short reading frames after annotation, since we have incorporated the criterion into the probabilistic framework. Although this still remains an arbitrary threshold which always imposes certain artificiality on our model, it is not an unreasonable measure to take, nor an uncommon one in the literature [Coffin *et al.*, 1997]. However, if we were not to have this rather *ad hoc* lower limit on gene lengths, it would not have a great impact on our annotation, but only result in slightly lower sensitivity and specificity

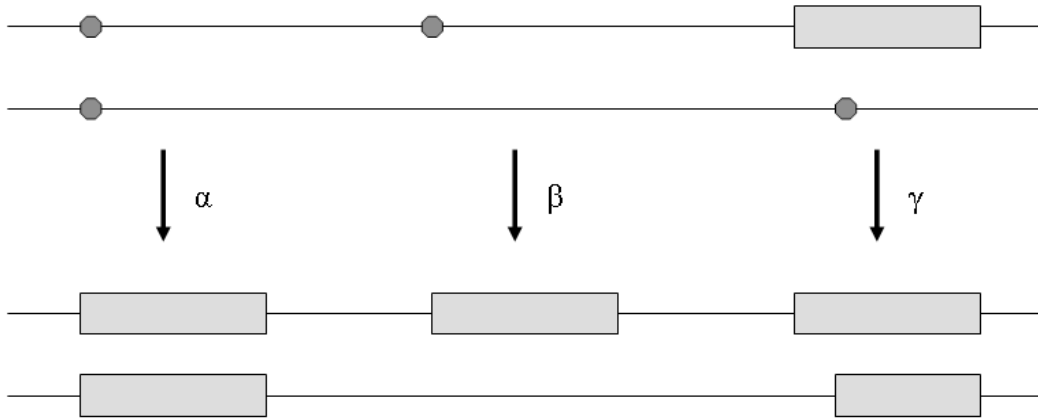


Figure 2.3: The 3 possible scenarios for entering a coding region in a particular reading frame, with their respective probabilities of transition  $\alpha$ ,  $\beta$ ,  $\gamma$ . An ATG is represented by a dot, a coding region in the particular reading frame by a box.

values of about 1%. Note also, that since we do not explicitly model gene length, any signal we get is purely due to our evolutionary model as opposed to marking open reading frames as coding merely because they are long. This is why we chose not to incorporate a prior on gene length into our model, in fear of this then being the driving force behind annotation of coding regions as opposed to evolutionary signal.

### 2.3.3 Emission Probabilities

Several models of varying complexity have been devised to describe the evolutionary substitution process between two nucleotide sequences, the most simple one being the Jukes-Cantor model [Jukes & Cantor, 1969]. Here a parameter  $g$  representing the estimated rate of evolution is introduced (see table 2.1), and assuming time  $\tau/2$  back to the most recent common ancestor of two sequences, which we may do due to the time reversibility of this model,

	A	C	G	T
A	$-3g$	$g$	$g$	$g$
C	$g$	$-3g$	$g$	$g$
G	$g$	$g$	$-3g$	$g$
T	$g$	$g$	$g$	$-3g$

Table 2.1: The Jukes-Cantor one-parameter model, where  $g$  is the rate of substitution from one nucleotide to another.

we will expect to observe  $g \cdot \tau$  substitutions per site between them. In reality transitions and transversions — transitions being substitutions of purine to purine and pyrimidine to pyrimidine, transversions respectively the opposite — occur at different rates though, the former generally occurring much more frequently.

This gives rise to the Kimura [1980] two-parameter model, which makes a distinction between substitutions occurring due to a transition and a transversion. As shown in table 2.2 we refer to the instantaneous rates at which these occur as  $g_{ts}$  and  $g_{tv}$  respectively. Assuming an evolutionary distance of  $\tau/2$  to the most recent common ancestor 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  $\tau$  individually. The probabilities  $X_{id}$ ,  $X_{ts}$  and  $X_{tv}$  of, after time  $\tau$ , at a certain locus observing an identity, transition and transversion respectively are therefore given by  $\exp \mathbf{Q}\tau$  where  $\mathbf{Q}$  is our instantaneous rate matrix in table 2.2:

	A	C	G	T
A	$-2g_{tv} - g_{ts}$	$g_{tv}$	$g_{ts}$	$g_{tv}$
C	$g_{tv}$	$-2g_{tv} - g_{ts}$	$g_{tv}$	$g_{ts}$
G	$g_{ts}$	$g_{tv}$	$-2g_{tv} - g_{ts}$	$g_{tv}$
T	$g_{tv}$	$g_{ts}$	$g_{tv}$	$-2g_{tv} - g_{ts}$

Table 2.2: The Kimura two-parameter model, where  $g_{ts}$  and  $g_{tv}$  are the transition and transversion rates and  $a = g_{ts} \cdot \tau$  and  $b = g_{tv} \cdot \tau$  are the expected number of transitions and transversions per site respectively where  $\tau/2$  is the time to the most recent common ancestor.

$$P_{id}(a, b) = 1/4 \cdot (1 + \exp(-4b) + 2 \exp(-2(a + b))) \quad (2.2)$$

$$P_{ts}(a, b) = 1/4 \cdot (1 + \exp(-4b) - 2 \exp(-2(a + b))) \quad (2.3)$$

$$P_{tv}(a, b) = 1/2 \cdot (1 + \exp(-4b)) \quad (2.4)$$

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. If we count the number of distinct amino acids arising when one alters each of the three nucleotide positions in each of the 16 potential contexts, we obtain table 2.4. From this we may 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 Table 2.4 shows that 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 (see Table 2.3). 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. Treating, for example, ATG as a type 1 site and ATA, ATC and ATT as type 4 sites, means however, that the approximations made by us are most likely to have minor implications.

We model the evolution of our sequences according to the Hein & Støvlbæk [1995] model, which is essentially an extension of Li *et al.* [1985]’s model 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. In a coding region in a particular reading frame this will 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’  $t = [t_1, t_2, t_3]$  — an array consisting of the degeneracy annotation of a nucleotide for each of the three reading frames. For a specific example of such an annotation array see figure 2.4.

Using this degeneracy annotation Hein and Støvlbæk incorporate the concept of selection factors into their framework: transitions and transversions

	T	C	A	G	
T	Phe	Ser	Thy	Cys	T
	Phe	Ser	Thy	Cys	C
	Leu	Ser	*	*	A
	Leu	Ser	*	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Table 2.3: The Genetic Code

xAA	1:1:1:1	xCA	1:1:1:1	xGA	2:1:1	xTA	2:1:1
xAC	1:1:1:1	xCC	1:1:1:1	xGC	1:1:1:1	xTC	1:1:1:1
xAG	1:1:1:1	xCG	1:1:1:1	xGG	2:1:1	xTG	2:1:1
xAT	1:1:1:1	xCT	1:1:1:1	xGT	1:1:1:1	xTT	1:1:1:1
AxA	1:1:1:1	CxA	1:1:1:1	GxA	1:1:1:1	TxA	2:1:1
AxC	1:1:1:1	CxC	1:1:1:1	GxC	1:1:1:1	TxC	1:1:1:1
AxG	1:1:1:1	CxG	1:1:1:1	GxG	1:1:1:1	TxG	1:1:1:1
AxT	1:1:1:1	CxT	1:1:1:1	GxT	1:1:1:1	TxT	1:1:1:1
AAx	2:2	CAx	2:2	GAx	2:2	TAx	2:2
ACx	4	CCx	4	GCx	4	TCx	4
AGx	2:2	CGx	4	GGx	4	TGx	2:1:1
ATx	3:1	CTx	4	GTx	4	TTx	2:2

Table 2.4: Type annotation table for a nucleotide in one reading frame depending on its context. (1:1:1:1) means one-fold degenerate, (2:2) means two-fold degenerate, (4) means four-fold degenerate. The other special cases may actually be classified as one of the three above depending on the context.

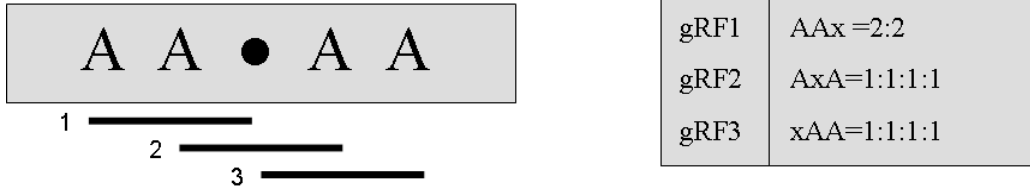


Figure 2.4: A nucleotide • in a certain given context. If this were a triple coding region the type-annotation array  $[t_1, t_2, t_3]$  would be  $[2, 1, 1]$ . For single coding in the third reading frame it would be  $[0, 0, 1]$ .

occur according to the Kimura model, and non-synonymous substitutions get accepted by a factor  $f$ . This means a mutation resulting in a change in amino acid is accepted by the factor  $f$ , compared to the case in which the amino acid has not been changed. Suppose a locus is of degeneracy  $[1, 0, 0]$ , i.e. the locus is only coding in GRF1 and a change in nucleotide would result in a change in amino acid regardless of whether it's a transition or a transversion. Thus our transition and transversion factors  $a$  and  $b$  get multiplied by our selection factor  $f$ . Now consider a site 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  $F_i([t_1, t_2, t_3])_{ts}$  and  $F_i([t_1, t_2, t_3])_{tv}$ . 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 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 an identity, transition and transversion after time  $\tau$  at a site of degeneracy  $[t1, t2, t3]$ , are given by

$$P_{id}(\tilde{a}, \tilde{b}) = 1/4 \cdot (1 + \exp(-4\tilde{b}) + 2 \exp(-2(\tilde{a} + \tilde{b}))) \quad (2.5)$$

$$P_{ts}(\tilde{a}, \tilde{b}) = 1/4 \cdot (1 + \exp(-4\tilde{b}) - 2 \exp(-2(\tilde{a} + \tilde{b}))) \quad (2.6)$$

$$P_{tv}(\tilde{a}, \tilde{b}) = 1/2 \cdot (1 + \exp(-4\tilde{b})) \quad (2.7)$$

where

$$\tilde{a} = a \cdot F([t1, t2, t3])_{ts} \quad (2.8)$$

$$\tilde{b} = b \cdot F([t1, t2, t3])_{tv} \quad (2.9)$$

with  $F$  as given in table 2.5. 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.

Ideally we would assign a different selection factor to each gene. However, since we are working in an HMM framework this is not possible. This is

however something we investigate further in Chapter 4 as well as the question of multiplicity of selection between overlapping genes. Another point is that start and stop codons may well be subject to more stringent selection than normal amino acid substitutions due to their special role in conserving gene structure. However, we will be dealing with genomes containing only a small number of genes, so again the amount of data available for a decent estimation of such factors would be too small. We therefore make no further assumptions about gene structure changes in our evolutionary model, especially since we allow for this via our transition probabilities.

Additional consideration needs to be given to regions where the gene structure has changed. In regions where gene structure differs, we can not 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.

### **2.3.4 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.

---

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

Table 2.5: 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  $f_{12}$ .

In the case of the transition probabilities we have 3 parameters:  $\alpha$ ,  $\beta$  and  $\gamma$ . 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  $\alpha$ ,  $\beta$  and  $\gamma$  occurred. For the case of  $\alpha$ , say, just group all expected transitions  $A_{ij}$  of type  $\alpha$  together and call this number  $E_\alpha$ . We also work out the expected number of times that the transition  $\alpha$  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  $\alpha$  is given by

$$\hat{\alpha} = \frac{E_\alpha}{E_\alpha + E_{1-\alpha}} \quad (2.10)$$

and similarly so for  $\beta$  and  $\gamma$ .

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 2.5, 2.6, 2.7), 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

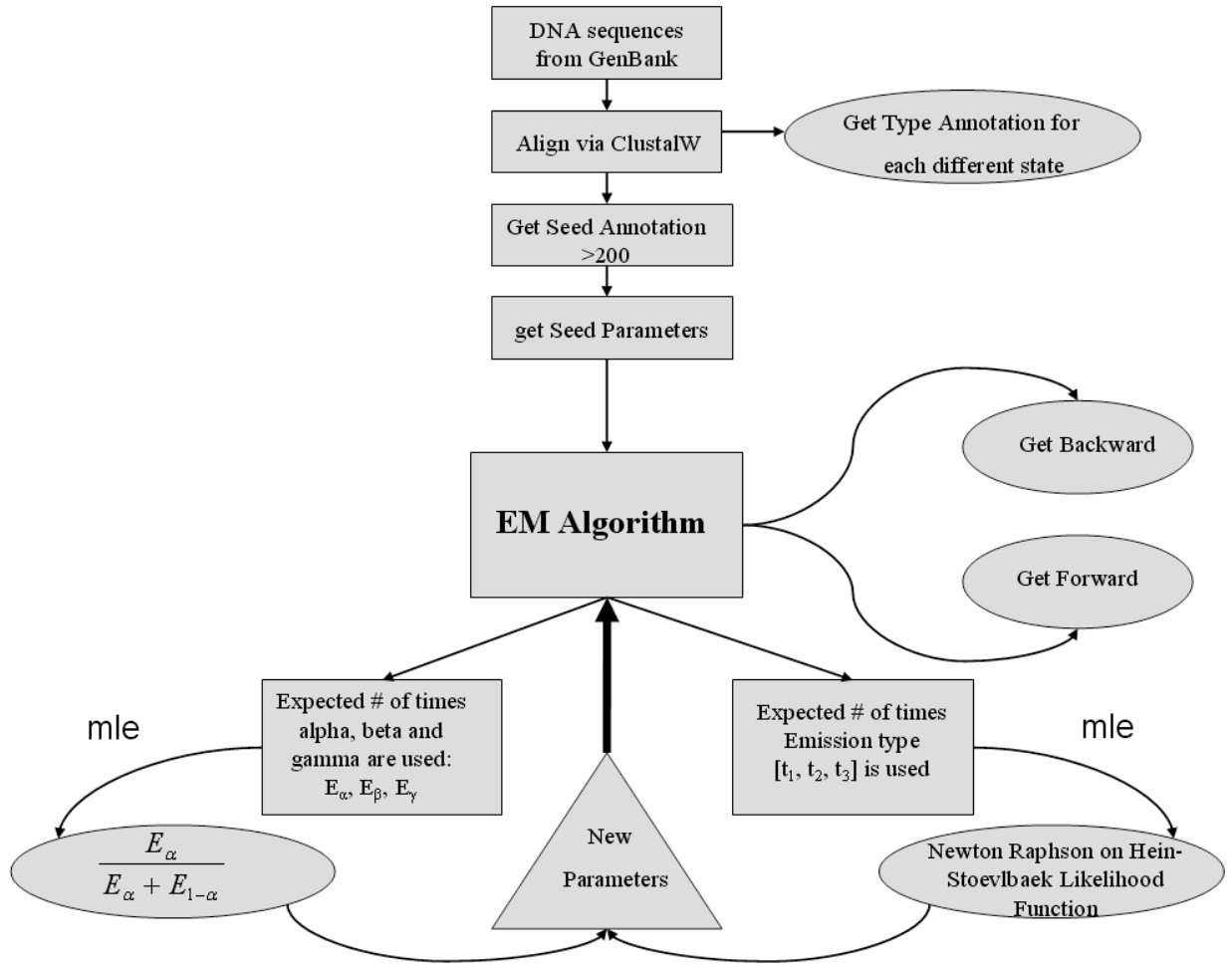


Figure 2.5: The structure of our annotative procedure

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].

### 2.3.5 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 & Hein [2006]. As true positives we take the sum  $\sum_i C^+(x_i)$  where  $x_i$  is the  $i^{\text{th}}$  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 =  $\frac{(TP-FN)}{TP}$
- Specificity =  $\frac{(TN-FP)}{TN}$

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

## 2.4 Results

### 2.4.1 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  $a + 2b = 0.1$  to 3.0. We then annotated the resulting alignment. Everything above a distance of  $\sim 0.2$  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  $\pm 0.035$  of their true value. However for more closely related sequences the quality of estimation for selection factors was much more volatile, deviating from the correct value by up to  $\pm 0.2$  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.

## 2.4.2 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

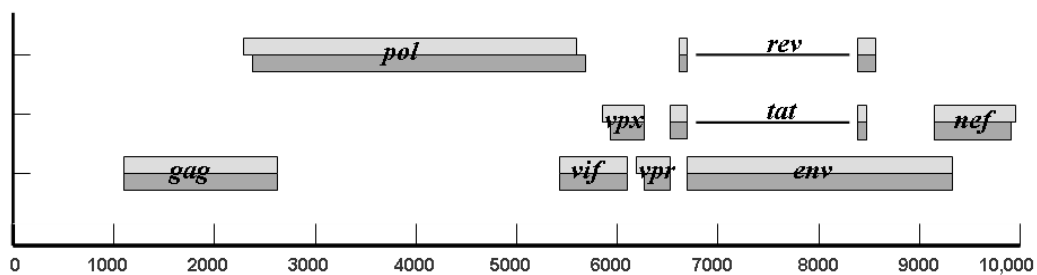
for this, which is generally a trivial exercise. We obtained the seed parameters for the EM algorithm by marking every open reading frame above 200 nucleotides as coding and subsequently calculated from this the maximum likelihood estimates of our parameters  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $a$ ,  $b$ ,  $f_1$ ,  $f_2$  and  $f_3$ .

### 2.4.3 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 2.6. We took a change of less than 1 in log likelihood as an indication of completed convergence and usually the EM algorithm converged sufficiently after about 3 iterations. A particular pairwise annotation of U27200 and J04542 is shown in figure 2.6, with the respective GenBank annotation above it. As one can see, our programme 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  $\sim 200$  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 purely on 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 2.4.7.

Table 2.6 and figure 2.7 show that we need sequences to have an appropri-

GenBank annotation of ClustalW alignment of strains J04542 and U27200



Pairwise HMM annotation of ClustalW alignment of strains J04542 and U27200

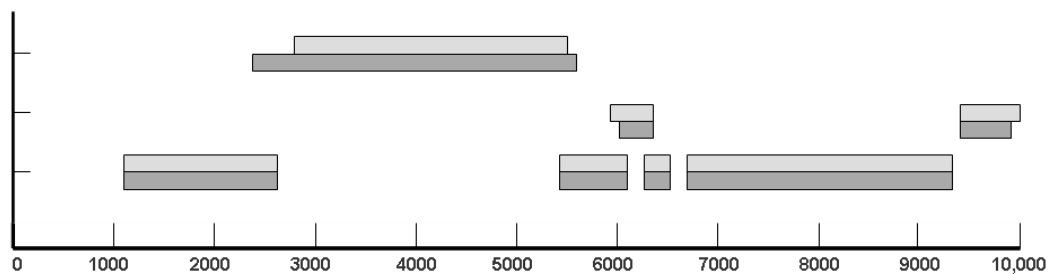


Figure 2.6: The annotation of HIV2 strands of U27200 and J04542. Above is the GenBank annotation and below the prediction of our programme. Each bar shows the genes in one sequence, with intronic regions being marked by single lines. Where a pair of bars does not overlap this indicates the change in gene structure via a shift in start or stop codon.

ate evolutionary distance to obtain any reasonable results, as our simulations already suggested. Figure 2.8 shows the phylogenetic tree given by MUSCLE [Edgar, 2004] underlying these six sequences and highlights the proximity of the pairs we have problems with — mainly M15390 and D00835, though it is unclear why we do so terribly badly on these. Looking at the estimate of selection factor  $f_2$  of 1.1 (which would imply an average positive selection on the genes in gRF2 — an unlikely scenario) it seems as though the sequences are actually even closer than the  $a$  and  $b$  values suggest, which would explain why the programme does so poorly. It is also reassuring to see that our predictions of evolutionary distances conform to the ones given in the tree. Using our comparative methodology it is thus unsurprising, that we see such bad results on the very closely related pairs. 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 belonging to either the *gag* or *pol* reading frame was unsurprisingly consistently slightly higher than the other two, due to the length of these genes.

Apart from that, one main drawback of the method, once used on sequences of sufficient distance, is a tendency towards suboptimal specificity (see figure 2.9). Due to generally finding some region of high conservation in a reading frame additional to one already coding, it tends to over annotate occasionally. Some of the loss in sensitivity is also due to ribosomal slippage having occurred, which we can not pick up on. Apart from that it does an excellent job in picking up ‘normal’ genes, i.e. ones without introns and with start and stop codons.

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

Table 2.6: 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 can not expect the predictions to be equivalent within one column. A graphical representation of this table is given in figure 2.7.

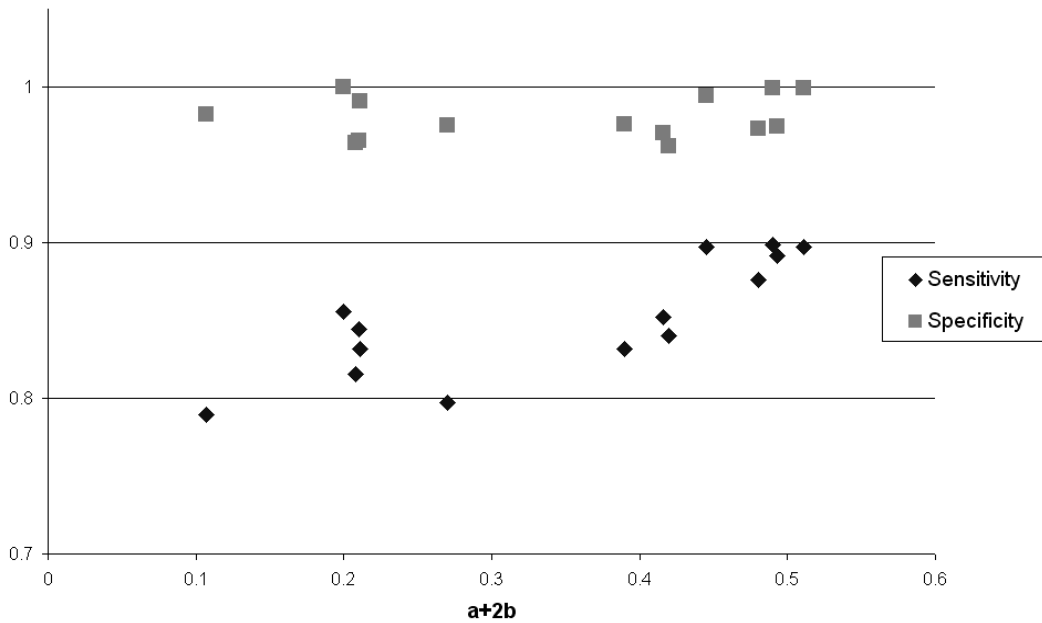


Figure 2.7: A graph representation of table 2.6 (without the M15390–D00835 comparison), with the evolutionary distance  $a + 2b$  along the  $x$ -axis, and sensitivity and specificity along the  $y$ -axis. We can see that with growing evolutionary distance our predictions tend to become better. Ribosomal slippage occurring in some strands also accounts for some of the fluctuation in prediction accuracy.

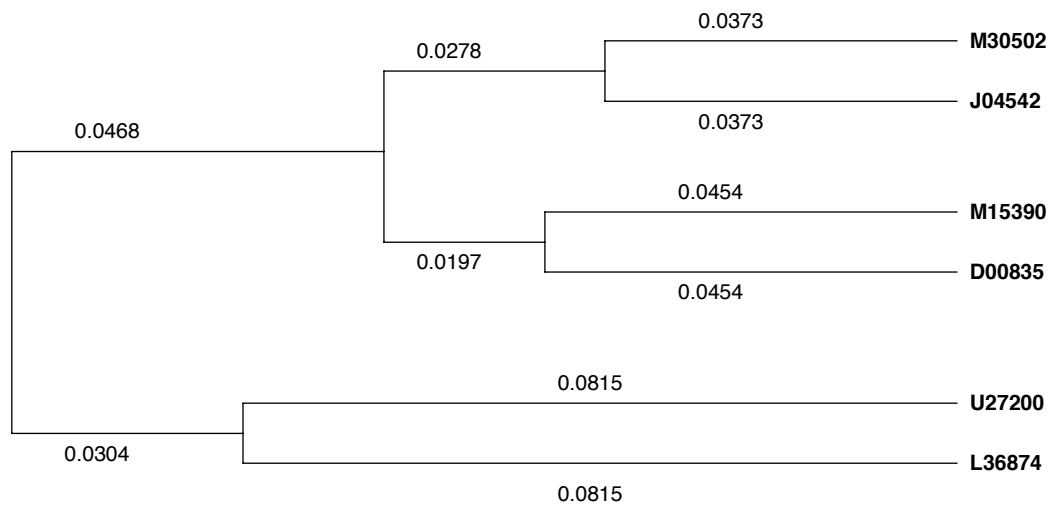


Figure 2.8: A phylogenetic tree relating the six sequences and their respective distances from the most recent common ancestor (from MUSCLE [Edgar, 2004]).

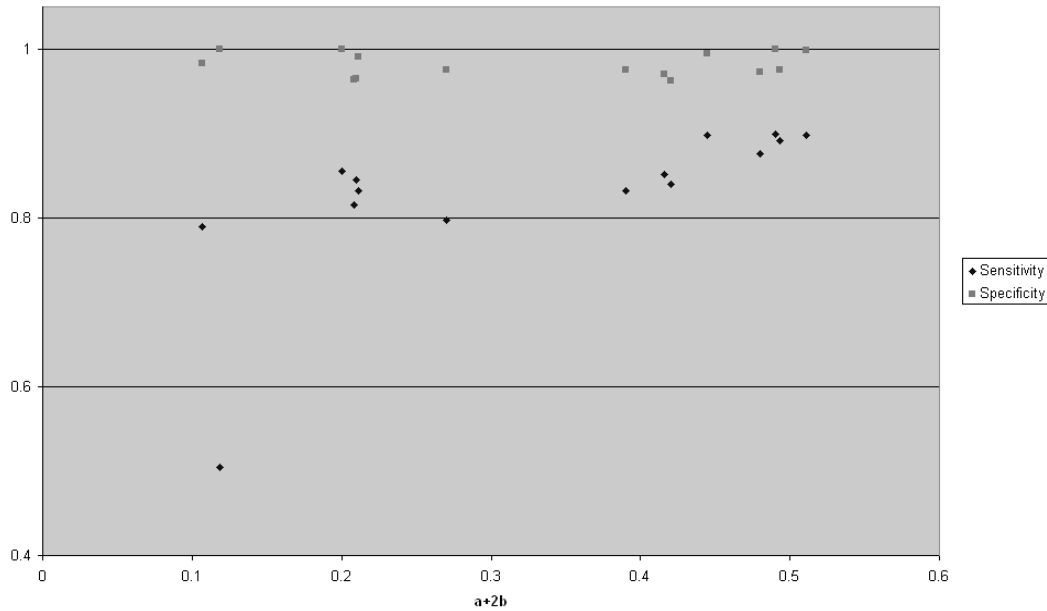


Figure 2.9: A plot of sensitivity and specificity of all pairwise comparisons run on the HIV2 genomes mentioned above in table 2.6 versus their evolutionary distance measured by  $a + 2b$

Another problem is that, although our programme is technically capable of picking up differences in gene structure whilst annotating both sequences, a comparative method will generally prefer to opt for a double coding state. It will only switch into single coding in a certain reading frame, if later on there is a single start codon in the other sequence in that reading frame (see figure 2.10) — basically if to achieve double coding it has no other option than entering coding regions separately in both sequences. In the case of a start codon having moved upstream, but the original ATG still being preserved it will generally be probabilistically preferable to start coding together in both sequences, due to our only considering evolutionary information.

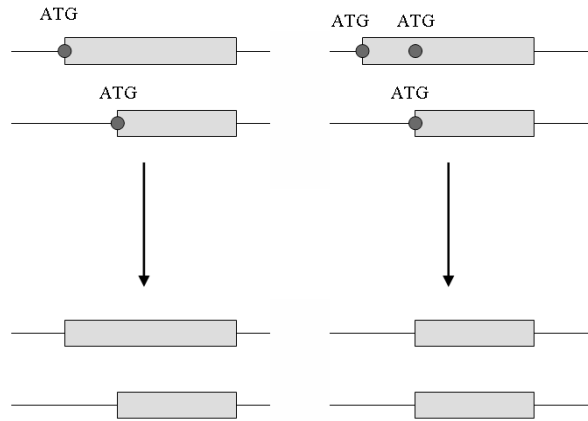


Figure 2.10: The real case scenario to the left, with genes being presented by black boxes and ATG codons by white spots. To the right, the structure prediction of our programme. This is naturally only true when  $\beta < \alpha$  which in reality is generally going to be the case.

#### 2.4.4 HIV1 vs. HIV2

We also ran the programme 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 programme GenAl introduced by Hein & Støvlbæk [1994, 1996], which combines DNA and protein 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 programme is dependent on an homologous region 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.

All in all, when comparing HIV1 and HIV2 we achieve an average sensitivity of about 80% and a specificity of about 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 can not 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 prob-

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

Table 2.7: 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.

lems mainly boil down to our annotation missing out on the *vif* and a few hundred nucleotides of the *env* gene. Looking at table 2.7 we can see the individual parameter estimates for these regions. *Gag* and *pol* will greatly 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.

## 2.4.5 Hepatitis B Virus

We applied our method to 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 100nt 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  $\sim 83\%$  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 2.11.

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%.

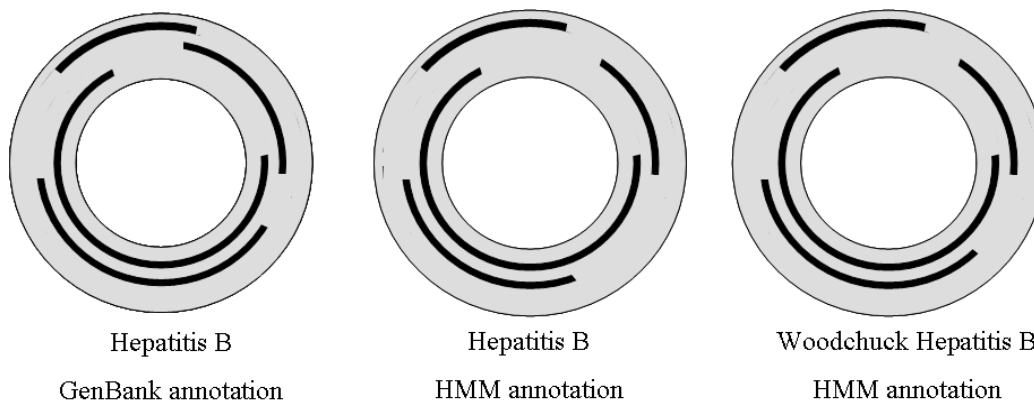


Figure 2.11: The circular Hepatitis B genome. To the left the GenBank annotation, which is nearly the same for both the Hepatitis B NC003977 and the Hepatitis B Woodchuck J02442 strands. To the right are the annotations for each strand predicted by our method. The dissimilarities in annotation arise due to our method choosing unaligned start codons further downstream than the true one.

## 2.4.6 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 is not the case. 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 figure 2.1). 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  $|\mathbf{I}_{k=s_i} - \epsilon|$ , where  $\mathbf{I}$  is the indicator function and  $\epsilon$  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 2.4.4 and in our Viterbi annotation fix the annotation of HIV1. For the sake of comparison we subsequently annotate HIV1 given HIV2. The results are close to perfect, achieving 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 and 94% specificity for the Woodchuck Hepatitis, respectively conditional on the other.

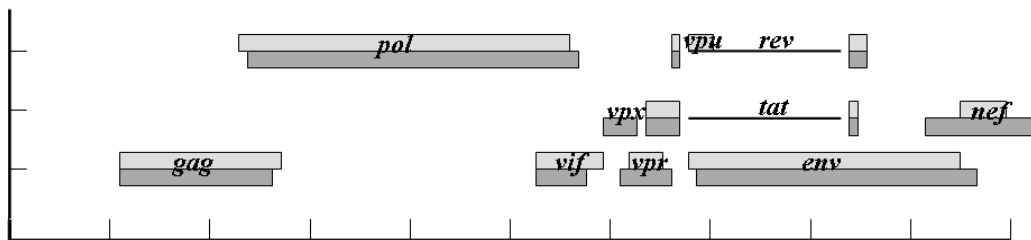
Although parameter estimates differ slightly, the final Viterbi annotation shown in figure 2.12 of the two sequences is close to identical, naturally apart from both times when the non-homologous genes in the other sequence are not picked up on. The only remarkable thing is the beginning of the *nef*

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  $\sim 50\%$  sequence similarity, the structural conservation over time provides us with enough information to annotate the homologous regions in the ‘unknown’ strand highly successfully.

### 2.4.7 Comparison to Other Methods

When comparing our results to other methods (see table 2.8), 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 for use on 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 allow for these. Similarly, GeneMark.hmm [Lukashin & Borodovsky, 1998] — used by Mills *et al.* [2003] to create the VIOLIN database —, which achieves comparable results to ours on the HIV

GenBank annotation of ClustalW alignment of strains HIV1 and HIV2



Pairwise Marginal HMM annotation of ClustalW alignment of strains HIV1 and HIV2

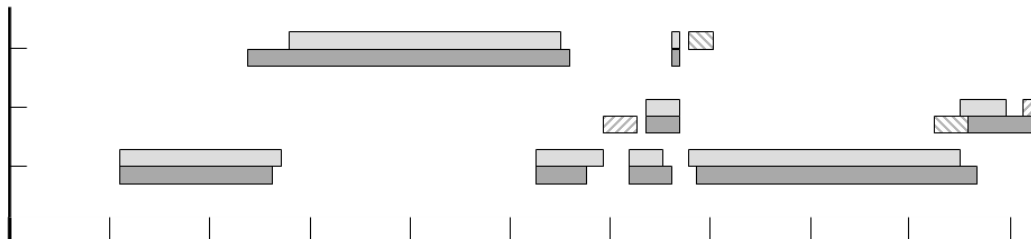


Figure 2.12: Above is the GenBank annotation and below the two marginal annotations of genomes K02013(HIV1) and M30502(HIV2). The dashed areas are the mispredicted areas in one of the marginal annotations, with left and right slanted dashes being annotated only when conditioning on HIV1 and HIV2 respectively.

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 for multiple coding regions, McCauley & 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 2.8 our performance is highly comparable even to the phylogenetic method, especially considering that our runtime is several orders of magnitude shorter.

Firth & 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 can 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.

---

Method	Overall Sens.	Overlapping Sens.	Overall Spec.
Firth & Brown	1.000	1.000	0.932
GeneMark.hmm	0.382	0.137	1.000
GLIMMER	0.589	0.286	0.974
McCauley & Hein Phylo	0.897	0.847	0.982
de Groot <i>et al.</i>	0.874	0.830	0.988

---

Table 2.8: 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.

## Chapter 3

# Annotation of Selection

## Strengths in Viral Genomes

*In the light of the prior chapter, we recognize the importance of dealing with the annotation of multiple sequences and the more realistic modelling of selection on viral genomes. We therefore continue by introducing a phylogenetic HMM which provides an ab initio annotation of a single multiple-coding viral genome, whilst modelling selection on a nucleotide basis. We thus show how to obtain both a structural and a selectional annotation of a given sequence. Our work is published in McCauley et al. [2007], and all programming was done by Dr. Stephen McCauley. Model design, simulation and experiment setup was done in collaboration with Dr. Thomas Mailund and me, and I am second author - and writer - of the above mentioned article.*

## 3.1 Abstract

### 3.1.1 Motivation

The coding of viral genomes in overlapping reading frames may result in atypical codon bias and particular evolutionary constraints. Due to the fast mutation rate of viruses, there is additional strong evidence for varying selection between intra- and intergenomic regions. The presence of multiple coding regions complicates the concept of  $K_a/K_s$  ratio, and thus begs for an alternative approach when investigating selection strengths. Building on the paper by McCauley & Hein [2006], we develop a method for annotating a viral genome coding in overlapping reading frames. We introduce an evolutionary model capable of accounting for varying levels of selection along the genome, and incorporate it into our prior single sequence HMM methodology, extending it now to a phylogenetic HMM. Given an alignment of several homologous viruses to a reference sequence, we may thus achieve an annotation both of coding regions as well as selection strengths, allowing us to investigate different selection patterns and hypotheses.

### 3.1.2 Results

We illustrate our method by applying it to a multiple alignment of four HIV2 sequences, as well as of three Hepatitis B sequences. We obtain an annotation of the coding regions, as well as a posterior probability for each site of the strength of selection acting on it. From this we may deduce the average posterior selection acting on the different genes. Whilst we are encouraged

to see in HIV2, that the known to be conserved genes *gag* and *pol* are indeed annotated as such, we also discover several sites of less stringent negative selection within the *env* gene. To the best of our knowledge, we are the first to subsequently provide a full selection annotation of the Hepatitis B genome by explicitly modelling the evolution within overlapping reading frames, and not relying on simple  $K_a/K_s$  ratios.

## 3.2 Introduction

The pressure on many viruses to minimize genome size often results in their coding in overlapping reading frames in order to compress information. The nucleotides involved in coding in such multiple coding regions tend to exhibit atypical behaviour, both in codon bias as well as in evolutionary constraints.

Various studies have been made of the evolutionary behaviour of overlapping regions. Generally, these regions have been thought to be under more stringent selection than singly coding sequences, since a mutation may result in a non-synonymous change in more than one gene. Indeed, higher than expected negative selection has even been used for the *de novo* detection of overlapping genes, as described by Spiropoulou & Nichol [1993] and Walewski *et al.* [2001].

Recently, however, several papers investigating overlapping coding regions — amongst them Guyader & Ducray [2002], Hughes & Hughes [2005], Mizokami *et al.* [1997], Osiowy *et al.* [2006], Pavesi [2006] — attempt to determine selection acting on overlapping reading frames by comparing non-synonymous to synonymous substitution rates. Within a pair of overlapping

genes, they report having detected purifying selection acting on one reading frame, and directional on the other. Unfortunately, using the concept of  $K_a/K_s$  ratio brings intrinsic problems with it, when applied to overlapping reading frames. Counting the ratio of non-synonymous to synonymous substitutions in one reading frame only makes sense, when the synonymous substitutions are unconstrained. In overlapping regions, however, a synonymous substitution in one reading frame could be non-synonymous in the other, thus biasing the analysis towards an underestimation of the ‘true’ synonymous substitution rate in one reading frame and thus the potentially false appearance of positive selection.

In this paper we introduce a model for the evolution of a sequence, which accounts for variability in selection along the genome. Each site may choose from a given set of selection strengths, ranging from directional to purifying, and sites are independent up to an auto correlation factor. We make no assumptions about the difference in selectional behaviour between overlapping and single coding regions, and thus give the model free rein. We incorporate our evolutionary model into the HMM introduced in McCauley & Hein [2006], extending it to a phylogenetic hidden Markov model, as described by Pedersen & Hein [2003] and Siepel & Haussler [2004]. Given a reference sequence and a multiple alignment, we use the phylogenetic HMM to annotate the former with protein coding regions, as well as different levels of selection — what we call a *selection annotation*. We so do not rely on a prior annotation off GenBank, thus making this method readily available for the coding and selection annotation of new virus strains. If however the annotation is available, we may use it as a prior and run our method solely

for the purpose of selection annotation.

## 3.3 Methods

### 3.3.1 Basic Structure of our Model

Initially, our set of coding states  $C$  is identical to those described in our previous paper [McCauley & Hein, 2006]. Since we are dealing with overlapping reading frames, each nucleotide may be coding in up to three different reading frames, and within each of these for three different codon positions. This results in there being 8 different coding states (see figure 3.1), on top of which we specify a set of  $m$  different selection strengths  $S$ . Now let  $X$  be an  $n \times L$  matrix, giving the gapped alignment of  $n$  sequences of length  $L$ , and  $T$  be the underlying phylogenetic tree, as shown in figure 3.2. Let the row vector  $X_1$  be our reference sequence and furthermore, let  $X_{k:l,j}$  denote the  $k^{th}$  to  $l^{th}$  entries of the  $j^{th}$  column vector. Let  $c = (c_j)$  ( $j = 1 \dots L$ ) be a vector of coding states — where  $c_j$  denotes the coding state at nucleotide position  $j$  — and similarly let  $s = (s_j)$  be a vector of selection states. Then we emit the joint probability distribution of an alignment column  $X_j$ , a coding state  $c_j$  and a selection state  $s_j$  with probability  $P(X_j, c_j, s_j)$ . Our model however is complicated by the fact, that both transition and emission probabilities are actually dependent on the nucleotides in their neighbourhood as well as their coding and selection states, as described in more detail in section 3.3.2:

- The transition probability from one coding state  $c_j$  to another is dependent on the prior three nucleotides.

- The emission of a nucleotide in the reference sequence is conditional on the prior two nucleotides.
- The emission of the alignment column vector  $X_{2:n,j}$  is conditional on the *nucleotide context*, by which we mean  $X_{1,j-2:j+2}$ .

Note that our model topology only allows us to transition from one coding state to another when it observes an emitted start or stop codon. Both the transition and the emission probabilities are thus dependent on the nucleotide context, as explained in more detail in section 3.3.2. The transition probabilities' dependence on the prior three nucleotides helps simplify the modelling of emission of a start and stop codon when respectively entering and exiting a new coding region. The emission probabilities' dependence on the nucleotide context allows us to capture both codon usage and is necessary for the identification of non-synonymous substitutions. The probability of observing a given column  $X_j$  together with an annotation  $(c_j, s_j)$  is then given by

$$\begin{aligned}
& P(X_j, c_j, s_j \mid c_{j-1}, s_{j-1}, X_{j-2:j-1}, X_{j+1:j+2}) \\
= & P(X_{1,j} \mid X_{1,j-2:j-1}, c_j) \\
& \cdot P(X_{2:n,j} \mid X_{1,j-2:j+2}, c_j, s_j) \\
& \cdot P(c_j \mid X_{1,j-3:j-1}, c_{j-1}) \cdot P(s_j \mid s_{j-1})
\end{aligned} \tag{3.1}$$

Effectively, what we are doing is splitting the emission of column  $X_j$  into the emission of the ancestral nucleotide  $X_{1,j}$  and its descendant column  $X_{2:n,j}$ .

In equation 3.1 therefore

- the first product term relates to the emission of the ancestral nucleotide conditional on the prior two nucleotides and the coding state,
- the second product term relates to the emission of the descendant column conditional on the ancestral nucleotide, its context and the coding and selection states
- the third product term relates to the probability of transitioning from one coding state to another dependent on the prior three nucleotides and finally,
- the fourth product term captures the probability of transitioning from one selection state to another.

Due to the various emission and transition probabilities having a two sided context dependency of up to three nucleotides, this holds for all but the first and last three positions of the sequence. These we fix to be in the non-coding state, which coincides with us wishing to annotate only ‘entire’ genes. In the case of being in the non-coding state, the column probability is independent of the neighbouring nucleotide context, i.e.

$P(X_j|c_i = NC, s_j, X_{1,j-2:j+2}) = P(X_j|c_j = NC, s_j)$ . We may then, using equation 3.1, formulate our likelihood  $L(X, c, s)$  of observing an alignment  $X$  together with a coding and selection state annotation, by simply multiplying the above expression over all columns  $j$ , ( $j = 4 \dots n - 3$ ).

Note, that while our model does not quite fit into the usual HMM framework, it is trivial to adapt the standard Forward, Backward and Viterbi

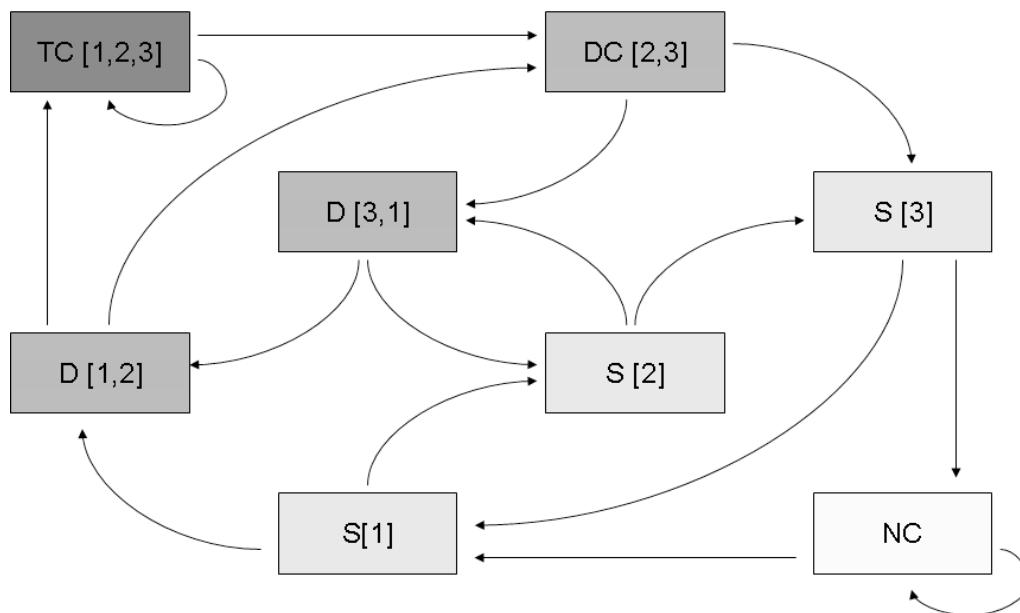


Figure 3.1: The topology sheet of the coding states in our HMM. Here NC stands for non-coding and S, D and T for single, double and triple coding respectively. The numbers pertain to the codon positions in each of the three potential reading frames. We may imagine the full HMM topology to be a layering of several sheets, one for each selection strength. The transition between coding states occurs within one sheet, whereas a transition in selection strengths occurs between two such sheets.

dynamic programming algorithms to it.

### 3.3.2 Model Parameters

We will now elaborate on how to calculate the factors of equation 3.1.

Our transition probability from a non-coding to a coding state is conditional on observing a start codon in the reference sequence. All other coding transition probabilities are deterministic — once in a coding state in a particular reading frame, one rotates in a three periodicity around the codon-positions until one meets a stop codon. This accounts for the coding state

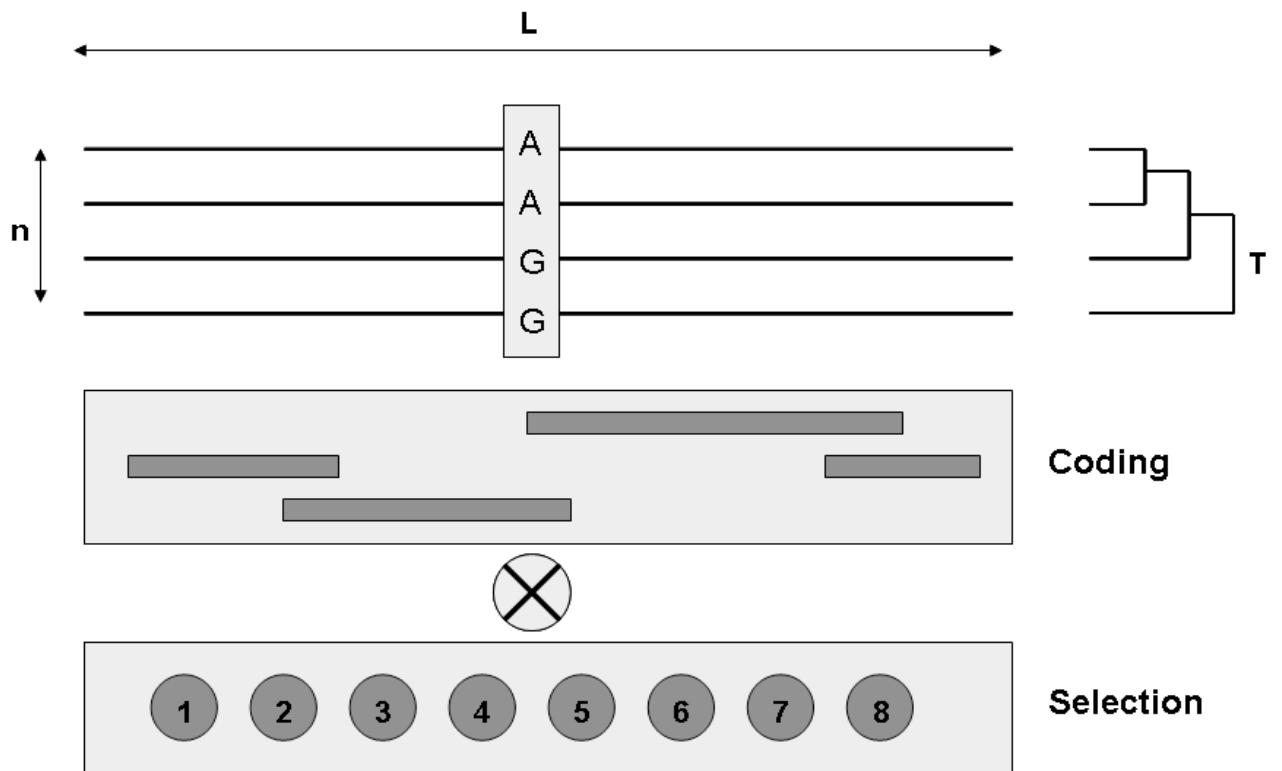


Figure 3.2: A graphical representation of our approach. Here we have  $n = 4$  aligned sequences of length  $L$  related by a tree topology  $T$ . We wish to annotate the reference sequence, by assigning both a coding state — given by one of the states in figure 3.1 — and one of the  $m = 8$  selection states. Thus the states in our HMM are effectively a product between the coding and the selection states.

transition probabilities being dependent on the previous three nucleotides in the reference sequence (see figure 3.1).

Concerning selection states, each one represents a different selection strength. These can be either independent parameters, or related by a grid to ease computational demands. Let our grid of size  $m$  be given by a fixed vector  $g = (g_1, \dots, g_m)$ . We then introduce a scaling parameter  $\omega$ , to obtain a new vector  $\omega \cdot g$  of  $m$  different selection strengths. The parameter  $\omega$  is merely an artificial tool to decrease the number of parameters, by having a single scaling parameter together with a pre-defined grid, and has no direct biological significance. The product  $\omega \cdot g$  is interpretable as the selection factor of a given site. Transitions from one selection state to itself occur with a certain probability  $\rho$ , which we call the auto-correlation factor, and states are switched with uniform probability  $\frac{1-\rho}{m-1}$ . With this model we cannot estimate single site selection, but only averages over subsets of sites. Without the auto-correlation, but keeping the grid of selection factors, we would essentially be modelling selection as a mixture model, similar to Yang & Swanson [2002]. The justification for including the auto-correlation is that we believe, that the structure of the underlying protein implies a dependency in selection of closely situated loci. With our model we will not pick up selective strengths of single nucleotide positions, but rather general trends in evolutionary behaviour of subdomains of proteins, such as recognizing important reactive sites.

The emission probability  $P(X_{1,j}|X_{1,j-2:j-1}, c_j)$  of the reference nucleotide in a certain coding state is dependent on the prior two nucleotides. It is drawn from independent multinomial distributions for each of the conditionals, thus

resulting in a total of 66 free parameters. For further details see McCauley & Hein [2006].

We let mutations in our sequences occur according to the general time reversible model (GTR), as described in Tavaré [1986], and obtain the base equilibrium frequencies from the nucleotide distribution of the reference sequence.

Given a reference sequence  $X_1$ , together with a set of aligned sequences  $X_i$  we now consider the effect a mutation at a certain position might have on the amino acid composition. Since we are dealing with overlapping reading frames, we must consider the reference nucleotide being — potentially simultaneously — at the first, second and third locus of a codon. The emission probability of the ‘descendant’ nucleotide is thus dependent on the nucleotide context in the reference sequence of two to the left and two to the right of it. Let  $E_{x,y}$  be the event of the mutation from nucleotide  $x$  to  $y$ , given a certain context  $X_{j-2:j+2}$  about  $x$ , causing *at least one* non-synonymous substitution. Notice that  $x$  is not necessarily the nucleotide  $X_{1,j}$  — the reference sequence is used to define the context of  $x$  only. With  $s_j$  being the selection strength on the sequence at position  $j$ , that mutation then gets accepted by a factor  $s_j^{I(E_{x,y})}$ , where  $I$  is the indicator function. If a substitution is synonymous this will assign it with a selection factor of 1. Additionally mutations which result in the loss or gain of a true stop codon get penalized by an additional factor  $s_{stop}$ .

We construct the substitution matrix  $A^j$  at column  $j$  as follows: Define  $F^j$  to be the matrix with entries  $s_j^{I(E_{x,y})}$ . That is to say, the  $(x,y)^{\text{th}}$  entry in  $F^j$  is  $s_j$  if and only if there is at least one non-synonymous substitution

between  $x$  and  $y$  in the context of  $X_{1,j-2:j+2}$ , else the entry equals 1. Let the rate matrix associated with our evolutionary process be  $Q$ . If  $\star$  denotes the entry by entry product between two matrices,  $(A \star B)_{kl} = A_{kl} \cdot B_{kl}$ , the instantaneous evolutionary rate matrix  $A^j$  for position  $j$  is then given by

$$A^j = \text{norm}(Q \star F^j) \tag{3.2}$$

where

$$\text{norm}(M)_{uv} = \begin{cases} M_{uv} & u \neq v \\ -\sum_{w \neq u} M_{uw} & u = v \end{cases} \tag{3.3}$$

We may subsequently obtain the expression  $P^j(t) = \exp(A^j t)$  for the substitution matrix of our model along a branch of length  $t$ , and from that calculate the probability of emitting a given column using Felsenstein's Pruning Algorithm on our tree topology  $T$ , with  $T$  re-rooted at  $X_1$ . Assuming column independence, this then gives us the likelihood of the data given the model. Given a set of seed parameters, we may thus estimate all free parameters using the EM-algorithm, combining Baum-Welch and numerical optimization. For the seed parameters, we first obtain a seed annotation by marking every open reading frame longer than 200nt as coding. From this we estimate our state-transition and evolutionary seed parameters, where initial input values to kick-off the optimization procedure make little difference to our final estimates.

One problem underlying our above approach is the fixed nucleotide context over time, and ideally we would let it vary. Pedersen & Jensen [2001]

investigated this by developing an exact evolutionary model, from which they sampled using Markov Chain Monte Carlo techniques, and observed markedly improved results in contrast to the stationary context model. However, as noted by them, the increase in computational complexity is unmanageable, and thus for now this slightly cruder version must suffice.

## 3.4 Results

To evaluate our method with respect to gene prediction, we use annotations from GenBank. We define true positives to be the sum  $\sum_j C^+(x_j)$  where  $x_j$  is the  $j^{\text{th}}$  nucleotide and  $C^+(x_j)$  is the number of reading frames it is coding in. Similarly we define the true negatives to be  $\sum_j C^-(x_j)$  where  $C^-(x_j)$  is the number of reading frames the nucleotide is not coding in. Then we may as usual define

- Sensitivity =  $\frac{(TP-FN)}{TP}$
- Specificity =  $\frac{(TN-FP)}{TN}$

With real genomes the ‘true’ selection annotation is not known, and we therefore use simulations to validate our method. As a measure of quality we use the average distance between estimated and true selection strengths over all coding sites.

### 3.4.1 Simulated Data

We initially wish to test our method on simulated data to see whether we can actually discern regions evolving under distinct selection as such. We

construct a 10,000 nucleotide long reference sequence containing one single coding region using the nucleotide composition of the single coding regions of the Hepatitis B genome. We subsequently fit a sinusoidal selection pattern along the genome with low and high peak values at 0.1 and 0.8 respectively and wave length of 250 nucleotides.

We evolve the reference genome accordingly along a tree into 3 descendant sequences, using the same evolutionary model as described in Methods. Subsequently we run our method to compare our selection annotation to the true one. As shown in figure 3.3, we are not annotating perfectly, especially as far as catching the peaks is concerned. However, although selection is rapidly changing along the genome, we still manage to capture the main essence of these changes and deliver a good representation of the true annotation. We present the selection annotation  $s_j$  for position  $j$  as a weighted posterior across all eight selection factors  $s_j = \sum_{m=1}^8 \omega \cdot g_m \cdot p_j(m)$ , where  $p_j(m)$  is the posterior probability of being in selection class  $m$  at position  $j$ , and  $g_m$  is the grid value for class  $m$  as described above.

Note also, that even on simulations with smaller wavelengths our model still recovers a good approximation to the true selection annotation.

We also wish to see what effect sequence divergence and length of consecutive coding regions have on our estimations. We simulate data with a differing percentage of average column identity across the alignment. In figure 3.4 we plot the column identity against our estimation error and observe the expected U-shaped plot, with errors for the evolutionary distances that we most likely would be dealing with being very low.

We also test our method on a double coding region of varying length

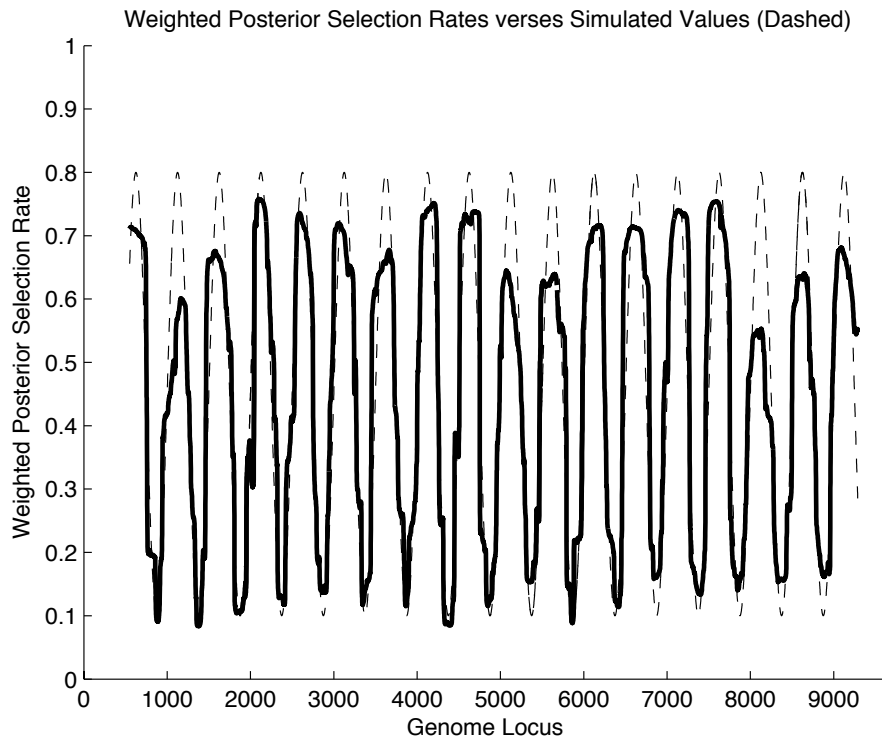


Figure 3.3: The simulated genome with the true sinusoidal selection given in dashed grey, and our weighted posterior annotation given in black.

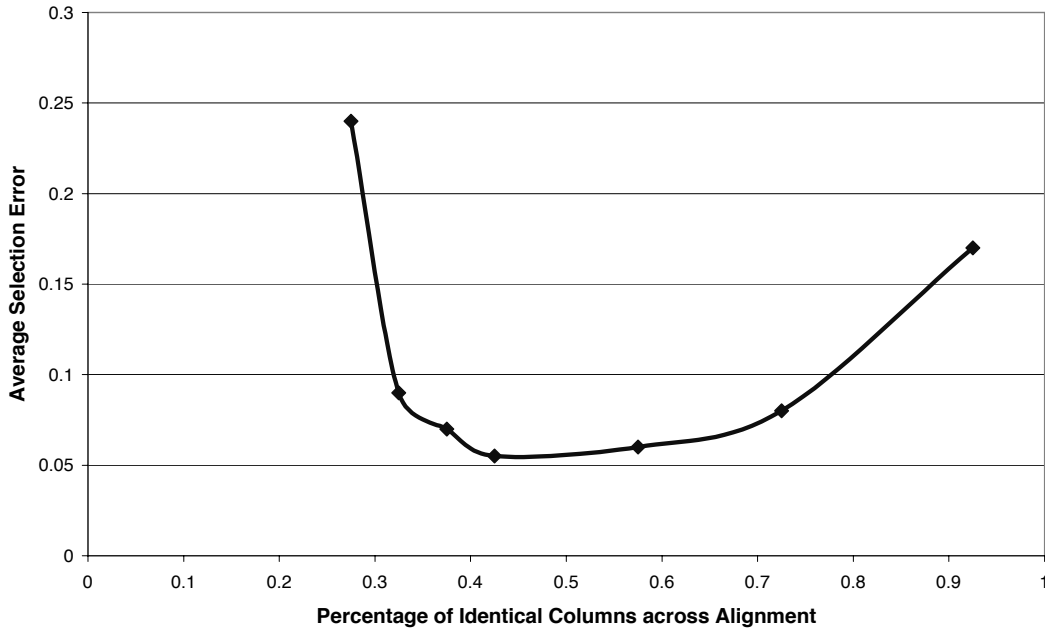


Figure 3.4: Simulation study showing the average percentage column identity over the alignment against our estimation accuracy. Both sequences too close and too far apart prove problematic to our method, since informational content decreases. Reasonable evolutionary distances however result in very good prediction accuracy.

acting under constant selection, to discern how long a region has to be for us to pick up a signal. Figure 3.5 shows that, as expected, very small lengths do not quite suffice for us to make out a signal, but we do very well, the longer the regions get. Since our method is not designed to pick up the selectional behaviour of individual nucleotides, but rather depict trends over genomic regions, this however is only to be expected.

### 3.4.2 HIV2

We illustrate our method on a small HIV2 dataset, where we annotate a reference sequence U27200 based on three descendant sequences AY530889, M30502, DQ307022 all of which are at a reasonable evolutionary distance to

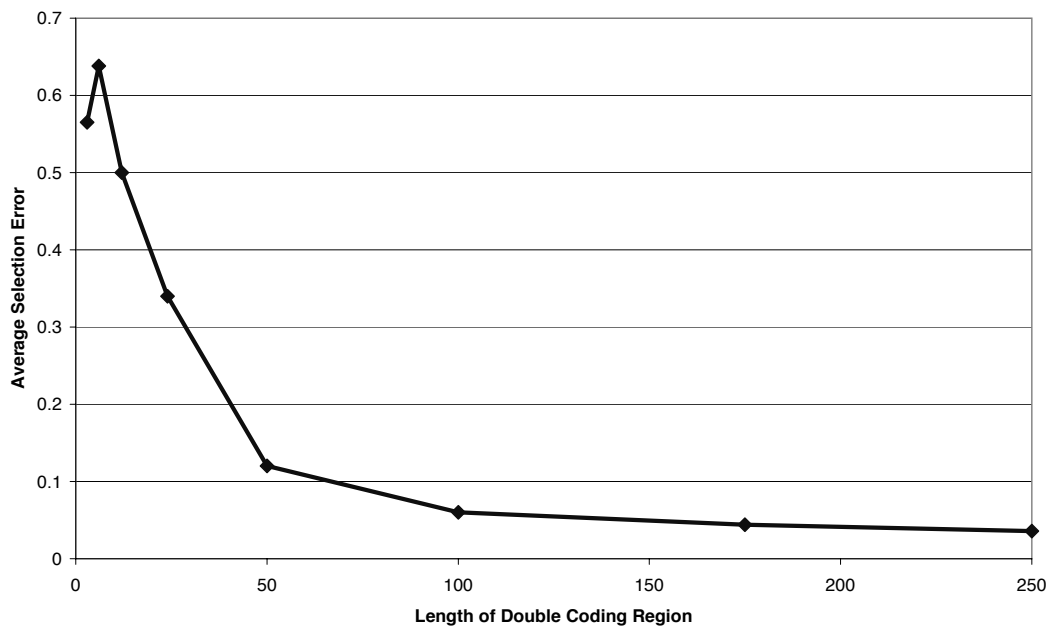


Figure 3.5: Simulation study showing the length of a double coding region, acting under constant selection of strength 0.8, against a background selection of 0.2 on a single coding region. The figure shows that below a length of 20 nucleotides our method has problems picking up the selection, but above that we rapidly increase in accuracy.

one another. To avoid complications due to recombination, we have chosen sequences which are believed not to be recombinants of one another, since unaccounted for recombination is known to bias the estimation towards the detection of positive selection [Scheffler *et al.*, 2006]. We download the sequences from GenBank and align them using CLUSTALW [Thompson *et al.*, 1994]. We subsequently obtain our phylogenetic tree using PHYLIP [Felsenstein, 1989]. The identity between the sequences ranges significantly from  $\sim 50\%$  to  $\sim 90\%$  across the genome. Due to CLUSTALW and PHYLIP's topology not accounting for the presence of overlapping reading frames, our use of them may of course be slightly problematic and we return to this in the Discussion. However, a goodness of fit test to the model implies that — apart from in the structurally conserved 5' and 3' UTR regions — our model fits the data well.

## **HIV2 Gene Annotation**

We first examine the gene annotation obtained. In comparison to the results presented in our earlier paper [McCauley & Hein, 2006], the inclusion of phylogenetic information improves slightly on the single sequence method, in particular by predicting the first half of the two intronic *rev* and *tat* genes. Since we require a start codon for the beginning of a coding region, our method is incapable of annotating the latter half of these intronic genes. Several small ORFs, both in the 5' and 3' UTR region, are falsely predicted as coding, presumably due to the secondary structure conservation in these regions. Our method uses no prior knowledge about the annotation of the descendant sequences. In the scenario, however, of wanting to annotate an

unknown strain with respect to a set of known ones, such knowledge is easily incorporated with a prior on the hidden state path.

As stated in McCauley & Hein [2006], the single sequence method alone already does a good job at gene prediction and improves on GeneMark.hmm, the state of the art method for viral genome prediction for single sequences [Besemer *et al.*, 1999, 2001, Mills *et al.*, 2003]. Once extended to a phylogenetic HMM, we note a marked improvement in our method achieving an average prediction sensitivity and specificity of  $\sim 97\%$ . This improves on the results in the recently published pairwise method presented by de Groot *et al.* [2007].

## HIV2 Selection Annotation

Next we consider the selection annotation. We choose to model our set of selection factors according to a grid of size 8 as shown in Table 3.4.2, scaled by the parameter  $\omega$ , see Methods. All selection parameter estimates are given in table 3.4.3. In this analysis the maximum likelihood estimate for  $\omega$  was 0.5512, scaling our grid to 0.0055–1.1024. Note here, that our model explicitly chose  $\omega$ , so that the majority of the states pertain to negative selection. Since it had the choice to accommodate for both positive and negative selection this is a marked decision, suggesting that overall selection is largely of a purifying nature.

The maximum likelihood estimate for the additional stop codon selection factor  $s_{stop}$  was 0.347, indicating that substitutions resulting in a change in stop codon are under three times as strong negative selection as normal non-synonymous substitutions. The autocorrelation parameter of transitioning

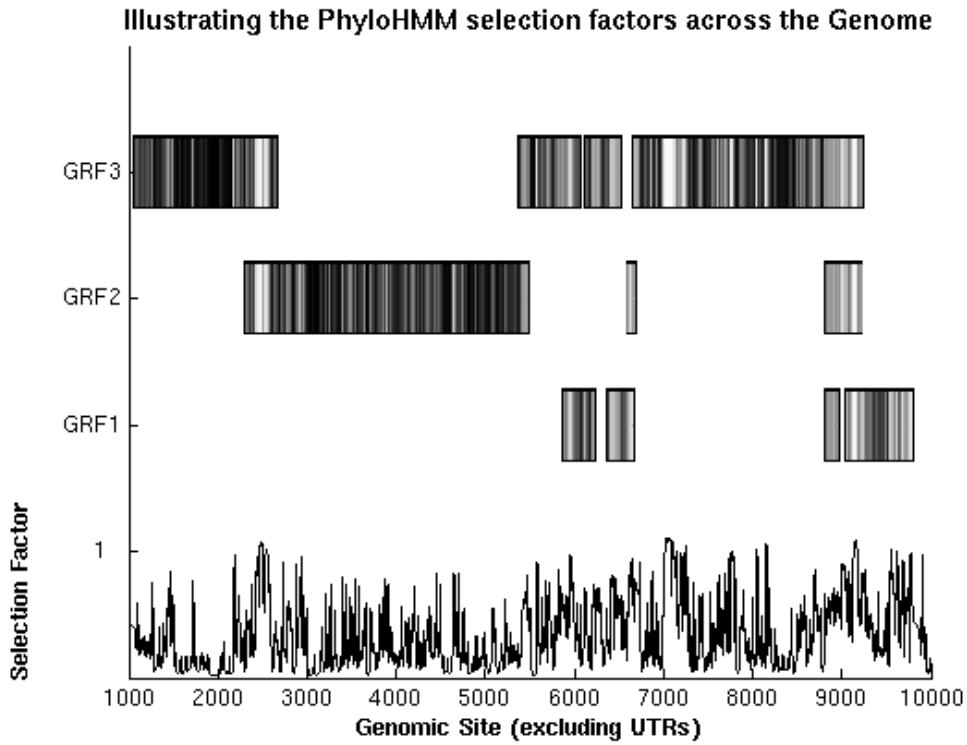


Figure 3.6: Selection Annotation across the HIV2 genome. In the above picture, the darkness of the shade refers to the strength of negative selection, with dark belonging to negative selection and the lighter regions closer to neutral. Below we give the selection annotation along the genome as the weighted posterior over the 8 selection states.

Selection Class	Step Factor	Factor $\cdot \omega$
1	0.01	0.0055
2	0.10	0.0551
3	0.25	0.1378
4	0.50	0.2756
5	0.75	0.4134
6	1.00	0.5512
7	1.50	0.8268
8	2.00	1.1024

Table 3.1: The grid of our selection factors, with the step factors in the middle fixed by us. The free parameter  $\omega = 0.5512$  is multiplied onto the steps to result in the final 8 selection strengths.

from one selection state to another was estimated at 0.932, suggesting a strong dependence between selection strengths at adjacent locations. This fits our expectation that neighbouring nucleotides are going to be under similar selection, since they will tend to belong to a common domain of tertiary structure.

Figure 3.6 illustrates the selection strengths along the genome. The darker the shading, the stronger the negative selection on that site. In the bottom part of the picture we present the selection annotation as given by the weighted posterior, as described in section 3.4.1. Interestingly enough, as shown in figure 3.7, the overlapping regions seem to be under less stringent selection than the single coding regions. In this figure, as in the following ones, the y-axis shows the average posterior probability. We obtain it by, for each of the eight states, averaging the posterior probability of being in that state over all nucleotides in the two types of region. Opinion on the

selectional behaviour of single vs. double coding regions is relatively split — Spiropoulou & Nichol [1993] and Walewski *et al.* [2001] assume constrained selection in double coding regions to identify *de novo* overlapping reading frames, whereas de Oliveira *et al.* [2004] seems to suggest more positive selection in double coding regions of HIV1 than in single coding ones. Since there is very little data in our analysis, however, we can not draw strong conclusions, but as figure 3.7 illustrates our method highlights less stringent selection in the overlapping regions of HIV2.

Figure 3.8 shows the weighted frequencies of the selection classes on the individual genes. The *gag* and *pol* genes are well documented to be under strong negative selection [Seibert *et al.*, 1995, Yang & Swanson, 2002], due to their housekeeping role in the viral organism. It is therefore reassuring to see that our results support this fact, with the selection strength distribution being biased towards the strongly purifying selection classes. The only surprise is maybe the fact that the *gag* gene is not under slightly more stringent conservation than in our observation. Our model can not pick up on exact sites of positive selection, but would rather be inclined to label a certain region under negative selection, which was sprinkled with many spots of positive selection, as being under closer to neutral selection. We find it thus unsurprising to see the other genes being under generally less stringent negative selection than *gag* and *pol*, due to the presence of precisely these sites.

One region in particular, around nucleotide position 7000, stands out in figure 3.6 as being very lightly shaded. Upon further analysis, this apparently highly divergent part of the *env* gene corresponds to a well-known hypervari-

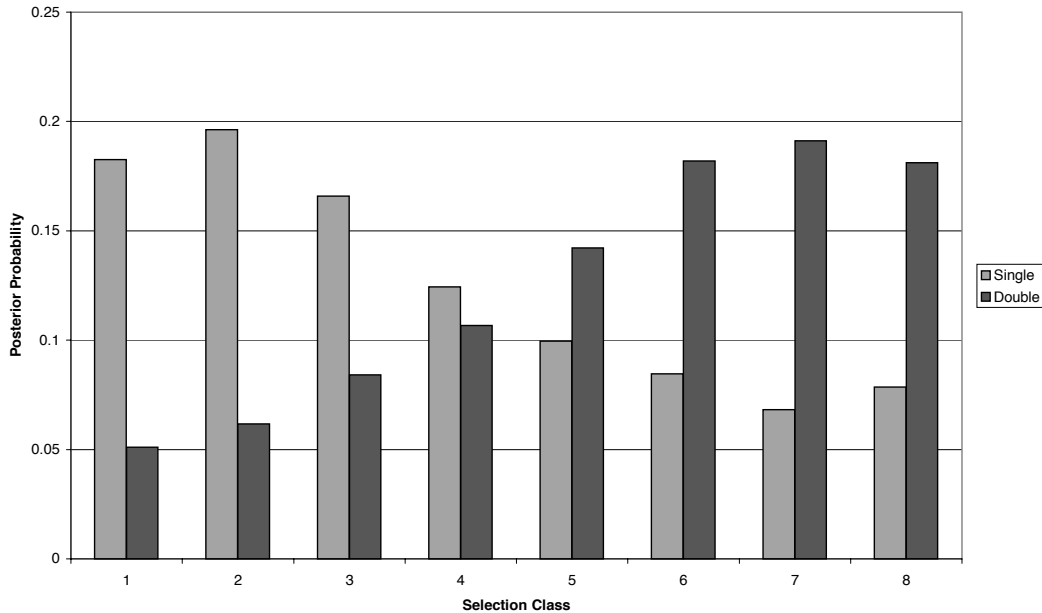


Figure 3.7: The distribution over the eight selection states for single and double coding regions in the HIV virus.

able region between two  $\beta$  sheets in the *gp120* protein, as cited by Simmonds *et al.* [1990], thus providing further indication of our method providing reasonable results. The *tat* and *rev* genes are very small intronic genes, implying that we have very little data on them to draw strong statistical conclusions. However, we do see a tendency towards more neutral selection in both of them, especially *rev*, such behaviour also having been documented by de Oliveira *et al.* [2004].

### 3.4.3 Hepatitis B

More than a third of the Hepatitis B genome is double coding, making it another excellent candidate to test our method. Several studies have been made on the selection within this virus, most recently by Osiowy *et al.* [2006]. However the fact that each one attempts to mold the concept of  $K_a/K_s$  ratio

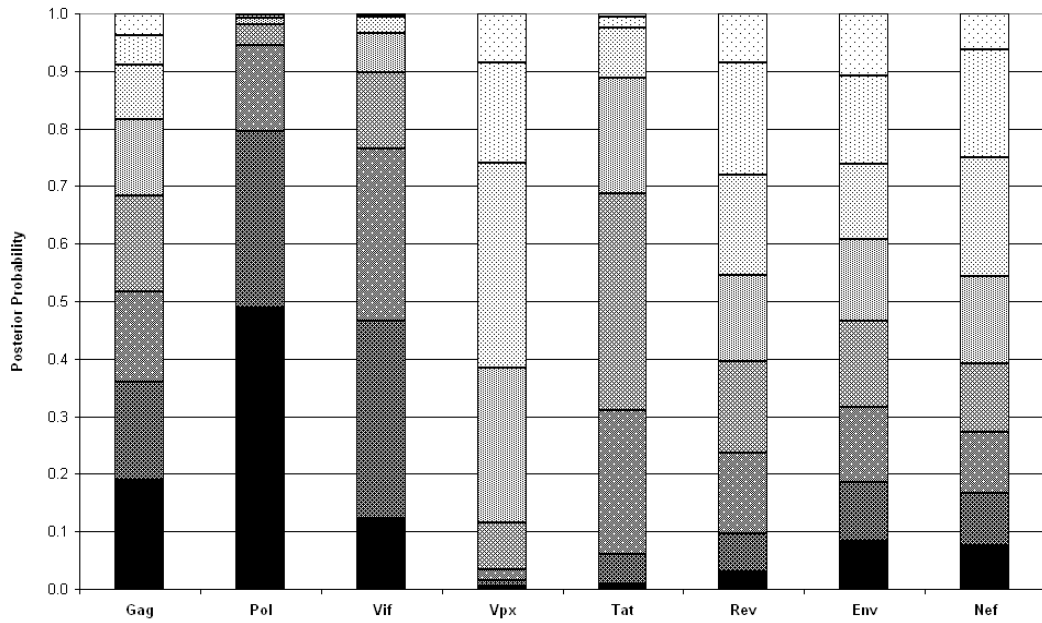


Figure 3.8: The selection strength distribution over all 9 HIV2 genes. For each gene we show the posterior probability of being in a given selection class. Here classes ranging from 1 to 8 are shown from bottom to top, with darker shades indicating stronger purifying selection.

to their purpose, provokes serious difficulties when attempting to analyse their results. We provide a full selection annotation of the Hepatitis B genome by explicitly modelling selection in overlapping reading frames, thus being to our knowledge the first to provide easily interpretable objective results on the evolutionary behaviour of this virus. We run our method on an alignment of three separate Hepatitis B strands X04615, M18752 and K02715.

### **Hepatitis B Gene Annotation**

Both the single sequence and the phylogenetic method predict the same gene annotation. We falsely predict a short overlapping ORF as coding as well as miss out on the first part of the S gene, resulting in an accuracy of 89.7% sensitivity and 98.2% specificity. Restricting our attention purely to the overlapping regions, we recover 84.7% of these. The complexity of the Hepatitis B virus particularly lends itself to highlighting the strengths of our method. For example GeneMark.hmm, used by Mills *et al.* [2003] to compile the VIOLIN database, only recovers 37% of the true coding regions and only 14% of the overlapping ones, demonstrating the true need for a more sophisticated approach. The pairwise comparative method by de Groot *et al.* [2007] achieves similar annotation results to ours, however does not provide a full selection annotation.

### **Hepatitis B Selection Annotation**

Figure 3.9 shows our selection annotation of the Hepatitis B genome. Due to the large amount of overlapping regions, it lends itself ideally to the study of selection acting on these. We can see that double coding regions are

on average under much stronger conservation than single coding ones, as demonstrated further in figure 3.10. Although, as shown in table 3.4.3, the grid factor  $\omega$  is estimated at 0.6 and the grid thus reaches from 0.006 to 1.2, there is effectively no posterior weight on the eighth rate class, and thus the entire Hepatitis B genome appears to be under negative selection. Interestingly, the separate selection factor for stop codons was estimated at 0.68 — nearly twice as high as in HIV2, suggesting that gene structure is less conserved in the Hepatitis B virus. Though, since we have a very limited data set this difference may well not be significant.

When looking at the selection strengths in figure 3.11 across the different Hepatitis B genes, we see that the C gene is certainly the most conserved, with 81% of the nucleotides coding for it being under extremely stringent negative selection of 0.06. There are a few light blocks within the genome, notably within the X and the C genes, potentially highlighting hypervariable regions within the Hepatitis B genome. Chain & Myers [2005] investigate the C gene using  $K_a/K_s$  ratio and find an annotation compatible with ours. Osiowy *et al.* [2006], however, conclude in their analysis of the X protein that it is under relatively stringent negative selection, thus seemingly contradicting our results, which quite clearly highlight a region of raised variability. Our method is not devised to explain, but merely to capture a tendency in selectional behaviour, so further investigation into the biological nature of these particular regions would be necessary. We do however return to the Hepatitis genome in section 4.4.2.

Genome	$\omega$	$\rho$	$s_{stop}$	$a$	$b$
HIV2	$0.55 \pm 0.03$	$0.93 \pm 0.01$	$0.35 \pm 0.04$	$6.23 \pm 0.16$	$2.05 \pm 0.06$
Hepatitis B	$0.60 \pm 0.05$	$0.98 \pm 0.01$	$0.68 \pm 0.08$	$11.20 \pm 0.55$	$6.97 \pm 0.30$

Table 3.2: The parameter estimates for the scaling grid factor  $\omega$ , the auto-correlation factor  $\rho$ , the selection factor  $s_{stop}$  on stop codons, the transition rate  $a$  and the transversion rate  $b$ , together with their variance. The branch lengths of the tree  $T$  were all estimated with very small standard errors.

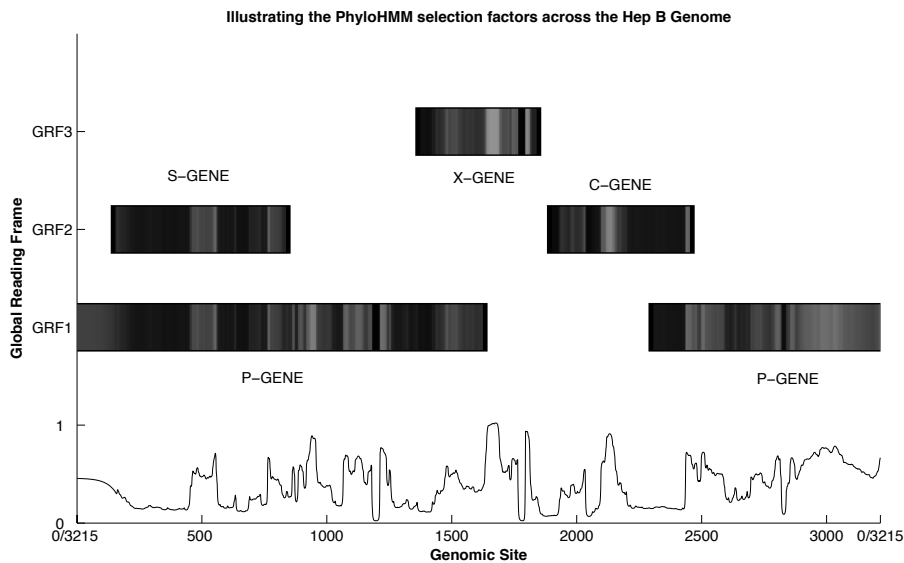


Figure 3.9: Selection annotation across the Hepatitis B genome. As before darker shades pertain to stronger negative selection. Below we give the selection annotation along the genome as the weighted posterior over the 8 selection states.

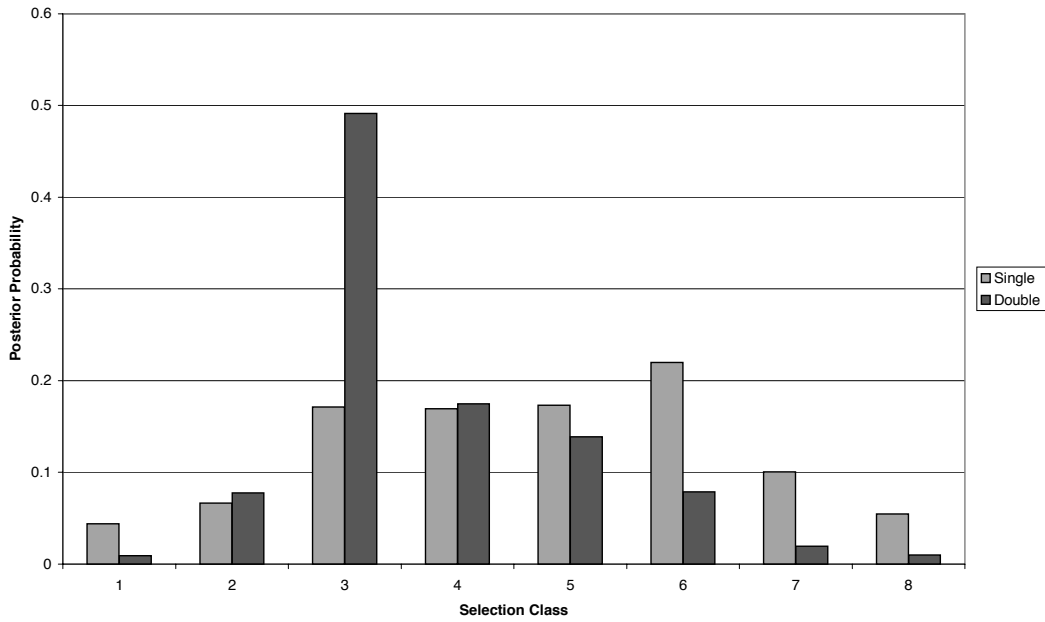


Figure 3.10: The distribution over the eight selection states for single and double coding regions in the Hepatitis B virus.

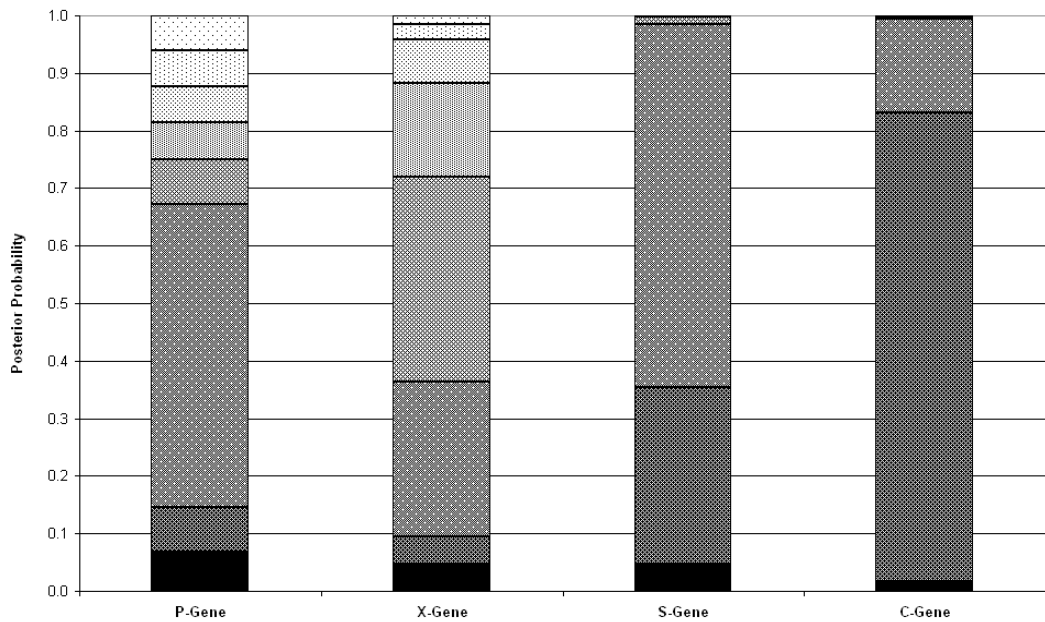


Figure 3.11: The distribution over the eight selection states for the four different Hepatitis B genes. For each gene we show the posterior probability of being in a given selection class. Here classes ranging from 1 to 8 are shown from bottom to top, with darker shades indicating stronger purifying selection.

## Chapter 4

# Investigating Selection: A Statistical Alignment Approach

*In order to further investigate selection on the viral genome, this chapter focuses on creating a counterbalance to modelling selection on a nucleotide basis. We devise a model which annotates selection for designated regions of the viral genome, and additionally wish to eliminate the bias created by the use of a fixed alignment. Our paper de Groot et al. [2007] is currently in review, and the model design was done in collaboration with Dr. Thomas Mailund. All programming, simulation and data analysis was performed by myself, except for the programming of the HMM compiler code, which was done by Dr. Gerton Lunter.*

## 4.1 Abstract

### 4.1.1 Motivation

We are interested in understanding the nature of selection on viruses, in particular Hepatitis B and HIV. Two problems complicate the study of selection in viral genomes: Firstly, the presence of genes in overlapping reading frames implies that selection in one reading frame can bias our estimates of neutral mutation rates in another reading frame. Secondly, the high mutation rate we are likely to encounter complicates the inference of a reliable alignment of genomes. To address these issues, we develop a model that explicitly models selection in overlapping reading frames. We then integrate this model into a statistical alignment framework, enabling us to estimate selection while explicitly dealing with the uncertainty of individual alignments. We show that in this way we obtain un-biased selection parameters for different genomic regions of interest, and greatly improve in accuracy compared to the fixed alignment method.

### 4.1.2 Results

We discover in HIV2 that double coding regions appear to be under less stringent selection than single coding ones. Additionally, there appears to be evidence for differential selection, where one overlapping reading frame is under positive and the other under negative selection. We also analyse Hepatitis B to understand the interaction of selection between two overlapping regions. We justify these observations with a series of simulation studies.

We show that the standard practice of fixing the alignment can lead to considerable biases, and that estimation accuracy increases substantially when explicitly integrating over the uncertainty in inferred alignments. We therefore propose that marginalizing over all alignments, as opposed to using a fixed one, should be considered in any parametric inference from divergent sequence data for which the alignments are not known with certainty.

## 4.2 Introduction

Popular belief has it that overlapping regions tend to be more constrained in their evolution than single coding ones, since a mutation may cause a non-synonymous substitution in up to three genes simultaneously. Indeed, this hypothesis has even been used for the *de novo* detection of overlapping genes, as described by Spiropoulou & Nichol [1993] and Walewski *et al.* [2001].

As stated before, various researchers have attempted to measure selection acting on overlapping reading frames, by investigating the  $K_a/K_s$  ratio within these regions for separate reading frames [Guyader & Ducray, 2002, Hughes & Hughes, 2005, Mizokami *et al.*, 1997, Osiowy *et al.*, 2006, Pavesi, 2006]. Comparing non-synonymous to synonymous substitution rates only makes sense when the synonymous substitutions are unconstrained. In the case of coding for multiple genes, however, a synonymous substitution in one gene may well be non-synonymous in the other and thus constrained. This will invariably bias the analysis towards an under-estimation of the ‘true’ synonymous substitution rate and thus lead to the false inference of positive selection.

Authors such as Rogozin *et al.* [2002] have attempted to resolve this problem, by focusing on synonymous substitutions in one reading frame which indeed are unconstrained in the other. Hein & Støvlbæk [1995] developed an evolutionary model particular to multiple coding regions, and used this for a study of selection on these. In chapter 2 we used this model of varying selection to comparatively annotate two viral genomes with evolved gene structure. In chapter 3 we incorporated a slightly extended version into their multiple sequence annotation method, which additionally provides a selection annotation of the genome. However, the last method looks at selection on an individual nucleotide level, and does not make assumptions about the modelling of selection on specific regions.

To study the imprint of evolution on viral genomes, it is necessary for the samples to have a reasonably high level of divergence. Since more divergent genomes are harder to align, this brings uncertainty about the alignment into the inference. We decide to circumvent this problem by considering the set of *all* possible alignments — and their corresponding likelihood under our model —, as opposed to a fixed ‘optimal’ alignment. This method has previously been used for similar purposes by Lunter *et al.* [2004] and Metzler *et al.* [2001], to minimize variability in parameter estimation due to uncertain alignments.

We work with a simple indel model, together with our evolutionary model, to generate a pairwise statistical alignment. For two sequences  $x$  and  $y$ , a set of seed parameters then gives us the probability  $p_{ij}$  of each  $i^{th}$  position  $x_i$  being aligned with each  $j^{th}$  position  $y_j$ . We subsequently work with expected observations as opposed to actual ones. We iteratively calculate the

alignment probabilities and the maximum likelihood estimates of evolutionary parameters, until we reach a given level of convergence. We also extend our methodology to a multiple pairwise method.

We run a simulation study to gauge the improvement made by considering all possible alignments as opposed to a single fixed one. We subsequently run our method on a set of 5 HIV2 sequences, as well as a set of 3 Hepatitis B genomes, and tackle various questions relating to overlapping reading frames and the selectional mechanism underlying these.

## 4.3 Methods

### 4.3.1 Outline

We describe the type of problem we are confronted with according to a specific example, shown in Figure 4.1. Due to the 3-periodicity of the genetic code, there are three global reading frames in which a sequence may code in the forward direction, henceforth referred to as GRF1, GRF2 and GRF3. In viruses these reading frames can tend to encode simultaneously for up to three different overlapping genes on each strand, resulting in multiple coding regions. We will be looking at single stranded RNA viruses, which predominantly code in the forward reading direction only. Amendments to our model would have to be made to include reverse reading frame encoded genes.

We are given two sequences  $S_1$  and  $S_2$ , descended from a common ancestor, together with the gene structure  $G$  of  $S_1$  — in the case of Figure 4.1 this is a genome with two genes which overlap. Say these genes code in

GRFs 1 and 2 respectively. Let us first assume we already have an alignment between our two sequences, and we wish to understand the way selection works on different regions of the genome. An initial question to ask would be, whether single and double coding regions behave in the same way. We thus, as shown, partition the genome into five segments, making a split wherever a gene starts or stops. These five segments we then assign to be of one of three region-types: non-, single- and double coding.

When considering the effect a mutation of the indicated nucleotide C in the overlapping region of  $S_1$  might have, we must consider its coding role in both reading frames. In GRF1 it is in the third position of the codon AGC and in GRF2 in the second position of the codon GCT. When looking at the genetic code, the codon AGx codes for serine or arginine, depending on whether x is a purine or a pyrimidine, respectively. On the other hand GxT codes for four different amino acids, depending on the nature of x. Therefore a transition in the nucleotide C will have no effect on the amino acid encoded by GRF1, whereas a transversion will. In GRF2 on the other hand, both will result in a non-synonymous substitution. Additionally, the selection strengths acting on either gene might be different, due to one of them evolving faster than the other.

Since we wish to analyse selection happening over a reasonable evolutionary distance, our aim is to be able to draw conclusions without relying on a prior alignment. Instead of estimating evolutionary parameters using *observed* substitution counts from a fixed alignment, we will therefore use an alignment model to generate *expected* substitution counts and from these use a maximum likelihood method to estimate all evolutionary parameters.

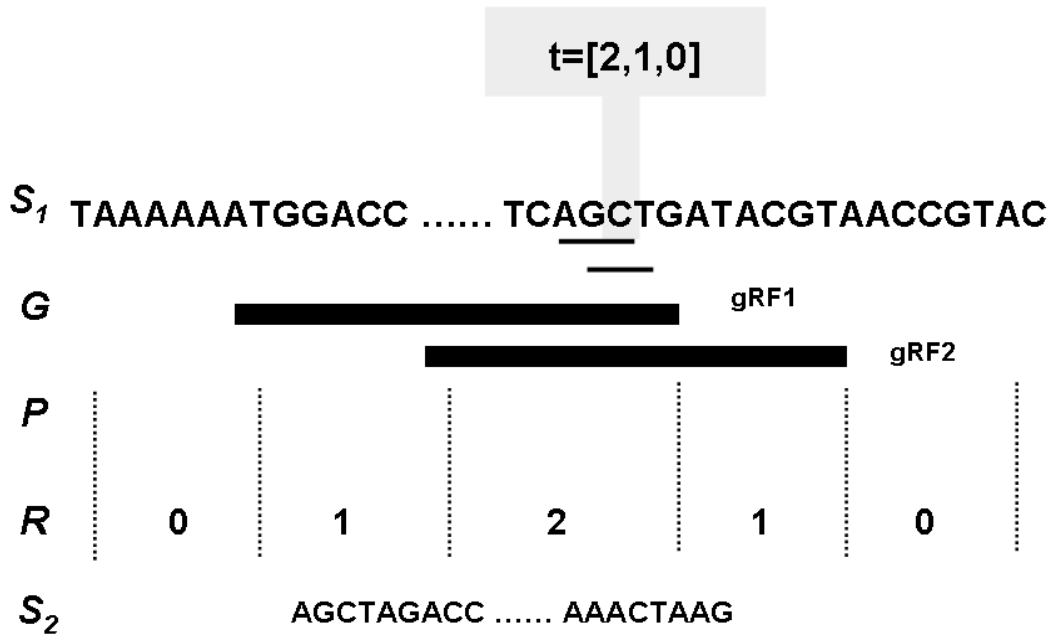


Figure 4.1: An example of our input data and annotation. We see here the ‘ancestral’ sequence  $S_1$ , whose genes structure  $G$  is given by coding regions in two reading frames GRF1 and GRF2. We apply a partition  $P$  to the sequence, where a breakpoint occurs whenever there is a change in gene structure. We annotate this partition with  $R = 3$  different types of region for non-, single and double-coding respectively. Finally we have the descendent sequence  $S_2$ .

In this manner we may sum over the uncertainty of the alignment — an uncertainty that will be high for distantly related viruses. Since our alignment model includes a substitution model, we iteratively switch between both it and our ML-procedure. Figure 4.2 depicts the basic outline of our programme.

### 4.3.2 Substitution model

To be able to calculate the probability of a certain alignment between  $S_1$  and  $S_2$ , we need to devise a model for the evolution of a sequence. We will be

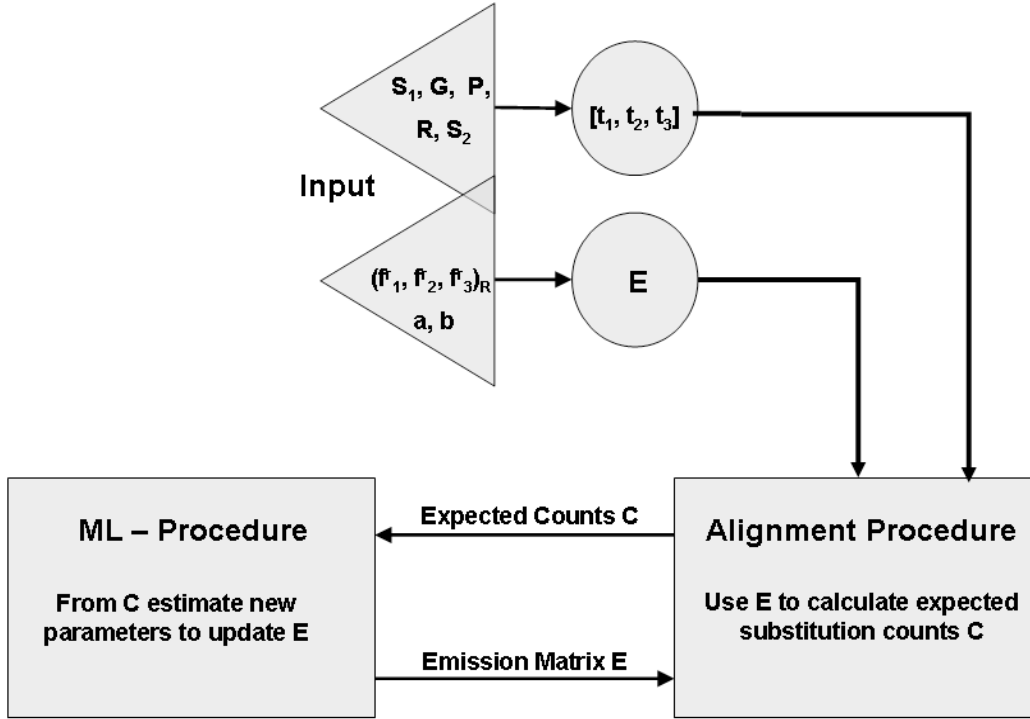


Figure 4.2: A graphic representation of our method. As input, we give the ‘ancestral’ sequence  $S_1$ , its gene structure  $G$ , our desired partition  $P$  and our region annotation  $R$  of the partition segments. We also input the ‘descendent’ sequence  $S_2$ , as well as our seed parameters for  $(f_1^r, f_2^r, f_3^r)_R$ ,  $a$ , and  $b$ . From this we may generate both our seed emission matrix  $E$  and the type-annotation-array  $t = [t_1, t_2, t_3]$  belonging to each locus along the sequence  $S_1$ . These then get input into our alignment procedure, which subsequently over the sum of all possible alignments, calculates the expected counts  $C$  of a certain substitution of a certain type in a certain region. This information gets transferred to our maximum-likelihood (ML) method, which generates our new parameter values, maximizing the expected observations  $C$ . The resulting emission matrix  $E$  gets fed back into our alignment procedure, and the loop continues until a change in parameters is below some given threshold.

working with a simple 3-state HMM indel model, using a more complex nucleotide substitution model, given by an emission matrix  $E$ , for the emission probabilities.

We wish to investigate region-specific selectional behaviour along the genome of  $S_1$ . We may thus apply a partition  $P$  to our sequence  $S_1$ , given by a sequence of partition points  $\{p_0, p_1, \dots, p_{|P|}\}$ , where clearly  $p_0 = 0$  and  $p_{|P|} = |S_1|$ . Because we are interested in certain global features, we may wish to group particular partition segments together into regions of a particular type. Say we have  $R$  regions, then each partition segment  $[p_k, p_{k+1}]$  with  $(0 \leq k < |P|)$  gets assigned to a certain ‘region-type’  $r$ , with  $(r \leq R)$ , where regions of the same type are assumed to evolve in a similar way.

As stated above, since we are interested in investigating the evolutionary behaviour of viruses in particular, we wish to work with a substitution model, which specifically accounts for the presence of multiple coding regions. For our evolutionary model  $E$  we use a model very similar to the one in de Groot *et al.* [2007].

As before, we model the evolution of our sequences according to the Hein & Støvlbæk [1995] model. That is to say, in our example in Figure 4.1 we can see an overlap between two genes, say genes  $A$  and  $B$ . This results in an annotation of  $[2, 1, 0]$  for our nucleotide  $C$  in the overlap, meaning that we have a degeneracy annotation of 2 and 1 with respect to gene  $A$  and  $B$  respectively.

Using this degeneracy annotation we incorporate the concept of selection factors into our framework: transitions and transversions occur according to the Kimura [1980] model, and non-synonymous substitutions get accepted

by a selection factor specified in the following.

Consider a nucleotide  $x$  in a region of type  $r$  in  $S_1$  with degeneracy-type array  $t$ . Then our factors will be given by  $F^r[t1, t2, t3]_{ts}$  and  $F^r[t1, t2, t3]_{tv}$ , for a transition and a transversion respectively. Within each region, we assign a selection factor to each gene within it, that is to say if gene  $A$  and gene  $B$  overlap in region  $r$ , we have selection factors  $f_A^r$  and  $f_B^r$  for mutations that result in a non-synonymous substitution in region  $r$  respectively in gene  $A$  and  $B$  only. In the case of a mutation causing a non-synonymous substitution in both genes, we would let it have selection factor  $f_{AB}^r$ . With our nucleotide of type  $[2, 1, 0]$ , this would mean that a transition would be multiplied by the selection factor  $F^2[2, 1, 0]_{ts} = f_B^2$ , since it would result in a non-synonymous substitution only in the amino acid in region 2 in gene  $B$ . A transversion however would be multiplied by the selection factor  $F^2[2, 1, 0]_{tv} = f_{AB}^2$  because it would cause a non-synonymous substitution in both gene  $A$  and gene  $B$ . If we were to assume independence between genes, the probability of a mutation causing a non-synonymous change in both genes would be given by  $f_{AB} = f_A \cdot f_B$ .

The probabilities of observing at a site of degeneracy  $[t1, t2, t3]$  in region  $r$  an identity, transition and transversion after time  $\tau$  are given by  $\exp \mathbf{Q}^r(t)\tau$  where  $\mathbf{Q}^r(t)$  is the appropriate instantaneous Kimura rate matrix:

$$P_{id}^r(\tilde{a}, \tilde{b}) = 1/4 \cdot (1 + \exp(-4\tilde{b}) + 2 \exp(-2(\tilde{a} + \tilde{b}))) \quad (4.1)$$

$$P_{ts}^r(\tilde{a}, \tilde{b}) = 1/4 \cdot (1 + \exp(-4\tilde{b}) - 2 \exp(-2(\tilde{a} + \tilde{b}))) \quad (4.2)$$

$$P_{tv}^r(\tilde{a}, \tilde{b}) = 1/2 \cdot (1 + \exp(-4\tilde{b})) \quad (4.3)$$

where

$$\tilde{a} = a \cdot F^r[t1, t2, t3]_{ts} \quad (4.4)$$

$$\tilde{b} = b \cdot F^r[t1, t2, t3]_{tv} \quad (4.5)$$

with  $F$  determined as explained above. We thus are able to construct an emission matrix  $E$ , where  $E(r, t1, t2, t3, x, y)$  is the probability of in region  $r$ , nucleotide  $x$  of type  $[t_1, t_2, t_3]$  mutating into nucleotide  $y$ .

### 4.3.3 Alignment model

We wish to eliminate the bias in parameter estimation created by the use of a fixed alignment. For this, we work with a *probabilistic* alignment, which instead of producing an 'optimal' alignment, computes posterior probabilities for each state at every nucleotide position.

To compute the probability of an alignment we use a simple indel model with Match, Insert and Delete states. We have as alignment parameters a gap-opening probability, a gap-extension probability and a transition probability from any state to the end state. All other state transition probabilities may be derived from these as shown in Figure 4.3. The Insert and Delete states emit a nucleotide from a uniform distribution, aligned to a gap. In the Match state nucleotide pairs are emitted according to our above model.

We thus will be considering all possible sequence alignments and weighing them appropriately (see Zuker [1991]), according to our indel model. Note, that when referring to the insertion and deletion states, the related posteriors are added together so that we obtain the posterior probability of a certain

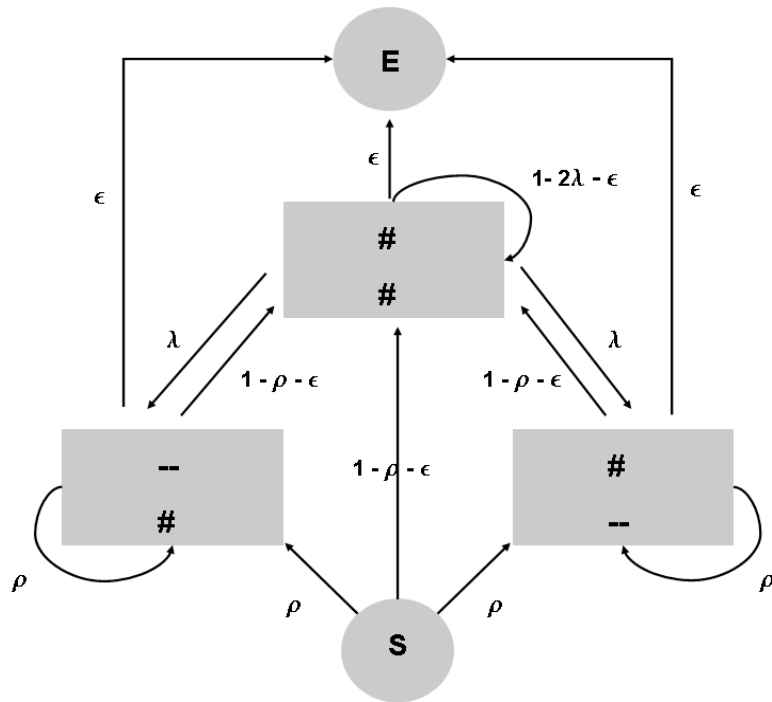


Figure 4.3: The states in our model, where  $\frac{\#}{\#}$  refers to Match-Match,  $\frac{\#}{-}$  to Match-Delete and  $\frac{-}{\#}$  to Match-Insert.

nucleotide not being aligned, as opposed to belonging to a particular gap.

During the alignment procedure, our alignment parameters are estimated in a few iterations of the Baum-Welch algorithm [Durbin *et al.*, 1998]. The implementation of the algorithm, including banding to cut computational demands, was generated automatically by the HMM compiler programme HMMoC [Lunter, 2007].

#### 4.3.4 Full model

As shown in Figure 4.2, we initially have as an input all the sequence and genome structure data, as well as a set of seed parameters. We subsequently use our alignment model to generate the posterior probabilities of every nucleotide position being in each state. From these we may calculate, for each degeneracy in each region, the expected number of times an identity, transition and transversion is used. For a site of degeneracy  $t = [t_1, t_2, t_3]$  in region  $r$ , let this be  $x_{id,t}^r$ ,  $x_{ts,t}^r$  and  $x_{tv,t}^r$  respectively. Since  $P_{id,t}^r$ ,  $P_{ts,t}^r$  and  $P_{tv,t}^r$  were the probabilities for a site of degeneracy  $t$  in region  $r$  of an identity, transition or transversion occurring (see equations 4.1, 4.2, 4.3), we may rewrite the emission term of the log likelihood as follows

$$\sum_t \sum_r x_{id,t}^r \log P_{id,t}^r + x_{ts,t}^r \log P_{ts,t}^r + x_{tv,t}^r \log P_{tv,t}^r$$

For this function of the  $3R + 2$  emission parameters  $(f_1^r, f_2^r, f_3^r)_R$ ,  $a$  and  $b$  we now find the maximum likelihood estimates using the Newton-Raphson iteration method and repeat the estimation step. Once the likelihood has converged, we generate a new emission matrix  $E$  to be fed back into our

alignment procedure in order to generate new posterior probabilities.

Once the likelihood function has converged below some set threshold, we output the final set of estimated selection parameters. We may also, if desired, construct an alignment — either using the Viterbi path, or posterior decoding — which in itself will be a superior alignment, to one not obtained via posterior weighting.

### 4.3.5 Extension to Multiple Sequences

We would like to be able to apply our method to multiple sequences, thus extrapolating more information where possible. We could of course devise a multiple alignment indel model, and develop a new likelihood function from which to maximize across all tree branches simultaneously. This however would be computationally much more demanding, runtime increasing exponentially with the addition of each new sequence. Instead, we therefore opt to work with a multiple pairwise alignment under the assumption of a rooted star shaped tree. This merely requires per additional sequence an extra transition and transversion parameter, since selection is acting on the gene in the ancestor and we assume this to be constant over all branches. The modification to our programme is thus trivial, with only a linear increase in runtime.

As an input we have, for lack of better terminology, the ancestral sequence  $A$  and its  $N$  descendants  $D_1, \dots, D_N$ , together with the seed parameters for the selection factors  $(f_1, f_2, f_3)_R$  on each region  $R$  as well as  $n$  transition and transversion parameters  $(a, b)_n$  respectively.

We then build a set of  $N$  pairwise alignments between the ancestor  $A$  and its  $N$  descendants. Each one of these obtains a likelihood function  $\log L$  as given in equation 4.3.4. Now we create a new likelihood function  $\log L^*$  which is the sum of the  $N$  log likelihoods. If  $x_{id,t}^{n,r}$  is the number of expected identities of type  $t$  in region  $r$  between the ancestral sequence  $A$  and its  $n^{th}$  descendant, then

$$\log L^* = \sum_n \sum_t \sum_r x_{id,t}^{n,r} \log P_{id,t}^{n,r} + x_{ts,t}^{n,r} \log P_{ts,t}^{n,r} + x_{tv,t}^{n,r} \log P_{tv,t}^{n,r} \quad (4.6)$$

is the full likelihood of observing all  $N$  sequences under our model. Note here, that the probabilities  $P$  are dependent on the sequence-dependent transition and transversion rates  $(a, b)_n$  and the selection factors  $(f_1, f_2, f_3)_R$  which in turn are *not* dependent on  $n$ , since we are assuming selection to occur on the gene in the ancestral sequence.

Maximizing this new log likelihood function, we proceed as above and estimate a new set of selection factors and a set of sequence specific transition and transversion rates, from which we may generate a new set of pairwise statistical alignments.

## 4.4 Results

### 4.4.1 Simulation

We test our method on simulated data, to see whether summing over all alignments does actually improve results notably. All the results in this section, unless stated otherwise, are obtained using the ‘worst-case-scenario’,

in the sense of having the least amount of data, of only two sequences.

By taking a 600 nucleotide sequence chunk out of a double coding region of the Hepatitis B NC00397 sequence, we construct a long double coding region, flanked by 300 nucleotides on either side of background sequence. We let this evolve according to the TKF91 model [Thorne *et al.*, 1991] into a descendent sequence, where the Match-Match state emits a descendant according to the Hein & Støvlbæk [1995] model with specified evolutionary parameters. We use a gap opening probability of 0.02 and a gap extension probability of 0.4 — these being values similar to the ones encountered in the real sequences we wish to analyse. We also only allow gaps of length 3 within coding regions, so as not to cause a frame shift in coding. We fix all selection parameters to 0.5 and test a variety of evolutionary distances, with transition rate  $a$  ranging from 0.2 to 0.7 and transversion rate  $b = a/2$ .

We annotate using our statistical alignment method described above, as well as performing parameter optimization on a fixed alignment produced by CLUSTALW. As we can see from Figure 4.4, we achieve better results on sequences closer together, but consistently outperform the fixed alignment scenario. Here performance is shown as the average absolute deviation of our estimated parameters to the true parameter values. The statistical alignment method performs, when applied to evolutionary distances we are realistically going to encounter, within 0.05 of the true value. Similar results hold for a number of other tested scenarios, including cases where one reading frame is under much stronger selection than the other and both are under positive or both under strong negative selection.

We can also see that using a fixed alignment causes a much more rapid

decrease in precision, than when summing over all possible alignments. This is unsurprising, since a greater evolutionary distance will produce a larger variation in possible ‘optimal’ alignments, with the chosen fixed one potentially being far away from the true one. For one, the greater the evolutionary distances get, the more likely it is for the CLUSTALW alignment to contain gaps of non-triplet length, thus creating a frame shift and subsequently completely off parameter estimates. Our statistical alignment method could be prone to similar shortcomings, but seems to overcome them by quite literally summing them away.

We wish to find out what effect the length of a double coding region has on our estimation accuracy. Letting the length of the double coding region in our above simulation vary from 600 down to 25, with transition and transversion rate 0.4 and 0.2 respectively, we obtain figure 4.5. As to be expected, the shorter the region, the worse our prediction results, since our data set decreases. However, above a length of 50 nucleotides we start picking up selection within a distance of  $\pm 0.15$ , and above 200 nucleotides we are within the  $\pm 0.1$  mark. We again consistently outperform the fixed alignment method.

We test the confidence levels of our predictions, trying to create as ‘realistic’ simulated data as possible. In the light of our real data analysis, we take the Hepatitis NC00397 genome and split it into 7 different regions, a new one starting whenever there is a change in gene structure. We evolve the sequence according to our indel model with varying transition and transversion rate of  $a = 0.2 - 0.8$  and  $b = a/2$  respectively, and fixed selection strength of 0.5 for each of the different regions. Depending on the evolutionary distance

and closely related to our results in Figure 4.4, we achieve an accuracy of approximately 70 – 94% with the statistical alignment method versus 20 – 72% for the fixed alignment method. Here our estimate is counted as correct if the true value lies within the error bars around the estimated value. The error bars around a parameter estimate are naturally highly dependent on the length of the coding region related to said error. They are however nearly identical for both the fixed and the statistical alignment method, and thus make the estimates comparable.

Gap placement in the ClustalW alignment often does not conserve the reading frame, and we expect that this is one reason for the comparatively bad performance of the fixed-alignment method. Manually adjusting alignments to conserve the reading frame indeed results in considerable improvement, thus demonstrating the volatility of results when dependent on one particular alignment. However, even when improving the fixed alignment, the resulting accuracy after manual adjustment still falls short of that achieved by the statistical alignment method, reaching only 40 – 70%.

Finally, we compare our results on the last setup using simulated descendants of the Hepatitis B genome in a pairwise versus a multiple sequence scenario. When adding up to four sequences, we observe the error bars getting notably tighter and simultaneously our estimation error decreasing by about 0.01 per added sequence. This implies, as desired, a more precise estimation of selection factors for multiple sequences.

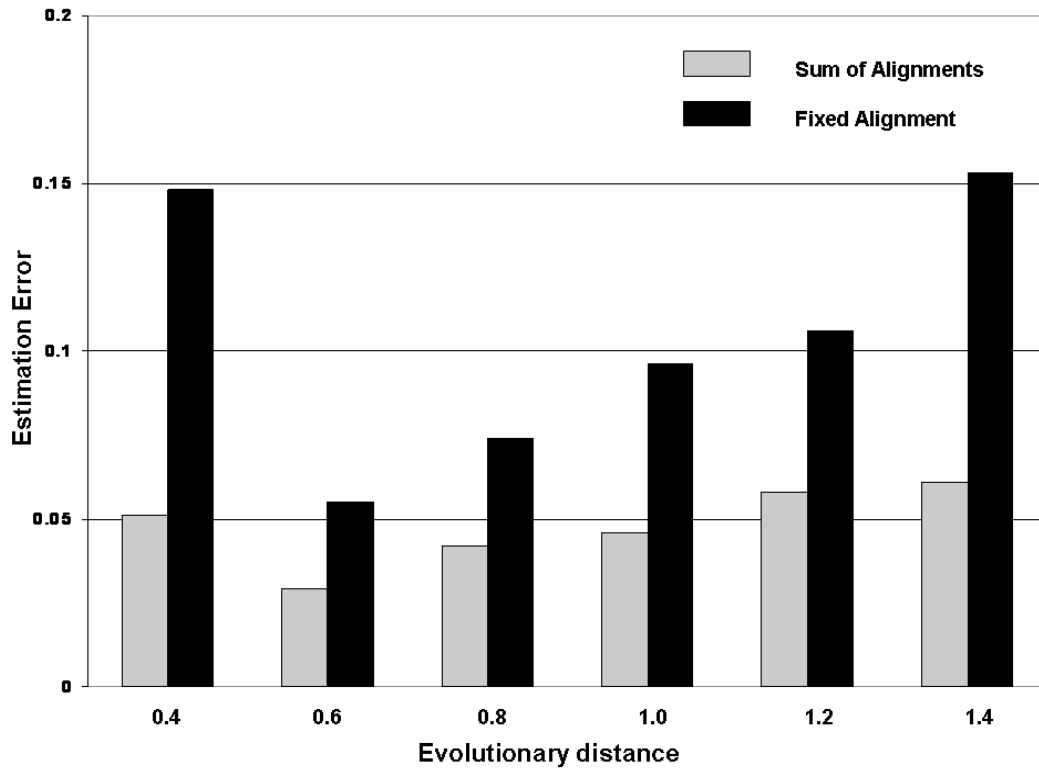


Figure 4.4: Simulation results between two sequences for a double coding region of length 600 of varying evolutionary distance. The figure plots the average estimation error of the statistical and the fixed alignment method, versus the evolutionary distance between the two sequences. The estimation error is measured as the average sum of the absolute deviations to the true parameter values of 0.5. The evolutionary distance is measured as  $a + 2b$ , where  $a$  and  $b$  are transition and transversion rates respectively.

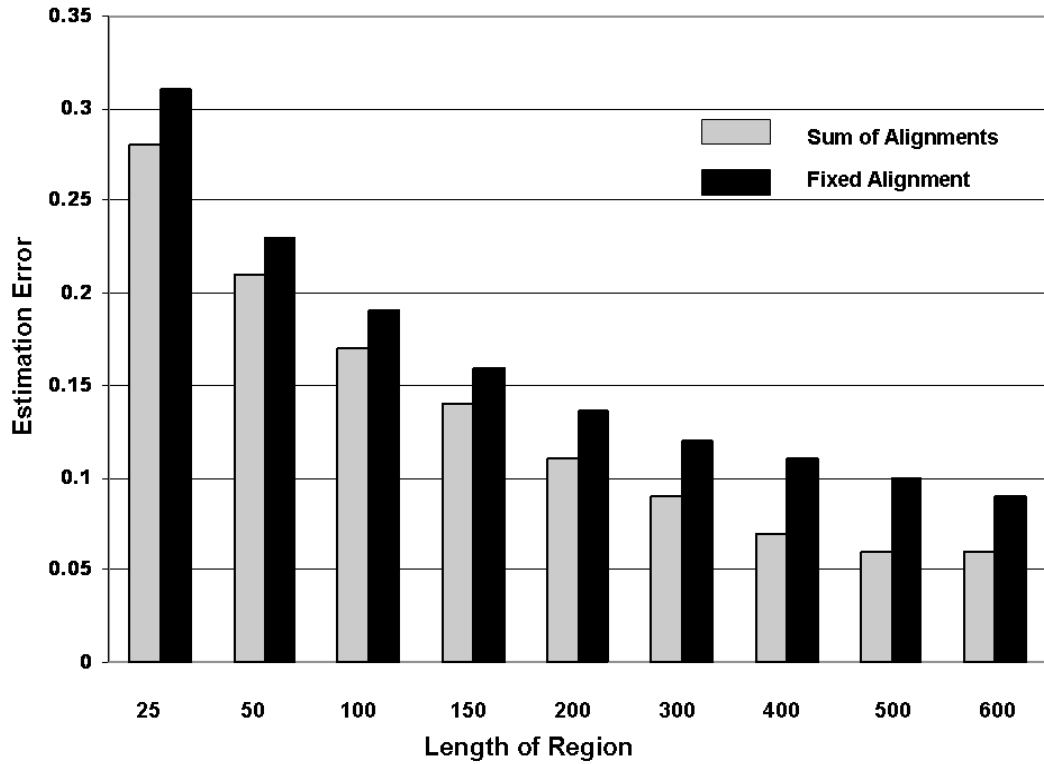


Figure 4.5: Simulation results between two sequences for a double coding region of varying length nested in a single coding region of length 800. The figure plots the average estimation error of the statistical and the fixed alignment method, versus the gene length of the double coding region. The estimation error is measured as the average sum of the absolute deviations to the true parameter values of 0.5.

### 4.4.2 Hepatitis B

We run our method on the Hepatitis B strand NC003977 and ‘descendants’ Woodchuck Hepatitis B strand J02442 and Ground Squirrel Hepatitis K02715, with sequences and gene structure downloaded from GenBank. As seed parameters we have all values set to 0.5 and wait between iterations for a difference in our loglikelihood of  $< 1$ . Our method takes  $\sim 40$  seconds to reach convergence and results are shown in Figure 4.6.

To see how a region acts when viewed as a whole, we also calculate the average selection acting on double coding regions, by weighting the expected counts for each mutation by the appropriate selection coefficient — in the case of a single non-synonymous change in gene  $A$  or  $B$  by the factor  $f_A$  and  $f_B$  respectively, and in the case of two non-synonymous changes by the joint factor  $f_{AB}$ . Table 4.4.2 shows the values obtained for the different regions, both single and double coding. We can see that when viewed like this, the double coding regions are on average under 0.41 selection, and thus not greatly different to the single coding ones at an average of 0.39.

Due to more than 1500 sites in the Hepatitis B genome being multiple coding, we may reasonably test whether the simpler multiplicative model is an equally good fit to the full one used above. Setting  $f_{AB} = f_A \cdot f_B$  we may perform a likelihood ratio test between the full and the restricted model, where selection acting on two different genes simultaneously gets multiplied up. With  $-2 \log \Lambda = 18$  for 3 added parameters, the full model fits the data significantly better than the restricted multiplicative one ( $P = 0.0004$ ).

Region	Type	Selection
1	Single	0.26
2	Double	0.31
3	Single	0.38
4	Double	0.40
5	Single	0.39
6	Double	0.46
7	Single	0.47

Table 4.1: The average selection acting on each of the seven regions of the Hepatitis B genome, measured by weighing each expected mutation by its appropriate selection coefficient. Selection on double coding regions appears to tend to be more lenient.

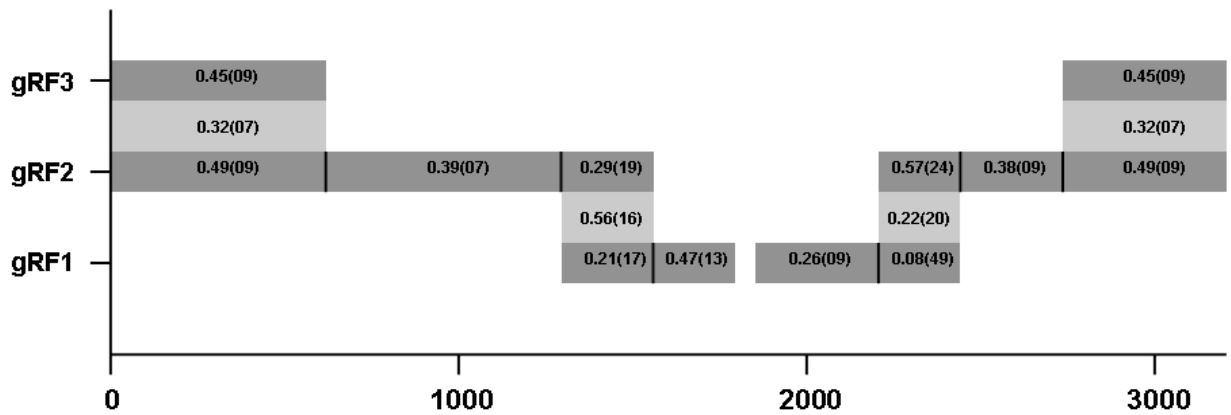


Figure 4.6: The estimated parameters for the seven different genomic regions of the Hepatitis B virus NC003977 based on the Woodchuck Hepatitis genome J02442 and the Ground Squirrel Hepatitis genome K02715. Here the darker shades refer to the selection acting on one gene only, and the lighter shades are the selection factors for non-synonymous substitutions in both genes. The error in the parameter estimates (one standard deviation) is given in brackets, in units of  $10^{-2}$ . The x-axis delineates the position on the genome.

### 4.4.3 HIV2

We apply our method to the HIV2 genomes J04542 with reasonably diverged ‘descendants’ U27200, M15390, DQ00835 and M30502, by splitting the genome into different regions whenever there is a change in gene structure. Setting all our initial parameters to 0.5, as above, we obtain a selection annotation for the different regions. The results of our parameter estimates are given in Figure 4.7.

As we can see, there is a marked difference between the estimated selection strengths underlying the different regions, with selection ranging from 0.21 – 1.50. Our results seem to suggest that genes encoded by double-coding regions often show contrasting modes of evolution, where one gene is highly conserved, whereas the other is less so. For example, in the second gene from the left, the *pol* gene, we see the middle section being under rather stringent selection of 0.24, whereas the two flanking regions are under less negative selection of 0.93 and 0.78. The respective overlapping sections in the other reading frames are under selection of 0.31 and 0.64. Similarly with the latter section of the following *vif* gene, we can see a dramatic increase in positive selection acting on the overlapping region, which rises to 1.50 against a selection of 0.23 in the other reading frame. Naturally all these estimates are made on relatively small regions, and thus have relatively large error bars, but tendencies towards a distinction between fast and slow evolving overlaps are nonetheless demonstrated. On the other hand, the selection on the overlap between the fifth and sixth gene in line — the *vpr* and the *tat* gene — is close to equal in both reading frames, thus indicating that the otherwise

observed high and low selection values are not mere inevitable artefacts of our model. We return to this in the discussion.

One of the most remarkable observations is that within each reading frame, selection on single coding regions appears to be *more* constrained than in double coding ones. As before, we calculate the selection acting on each region as a whole, as shown in Table 4.4.3 and see that on average the single coding regions are under selection of strength 0.39 whereas the double coding regions seem to be under less stringent selection of an average of 0.64. This is in line with the results shown in de Oliveira *et al.* [2004] and more recently in McCauley *et al.* [2007], but somewhat contrary to general belief [Spiropoulou & Nichol, 1993, Walewski *et al.*, 2001]. Clearly within the HIV2 genome there is much less data than with Hepatitis B, so it is harder to assign a true significance to these figures. However, our results do appear to suggest less stringent selection on overlapping regions than on single coding ones, thus maybe indicating the overlapping regions to be a relatively young feature in the virus.

Region	Type	Selection
1	Single	0.21
2	Double	0.51
3	Single	0.24
4	Double	0.54
5	Single	0.36
6	Double	0.63
7	Single	0.28
8	Single	0.60
9	Double	0.44
10	Single	0.45
11	Double	0.70
12	Single	0.45
13	Double	0.99
14	Single	0.57

Table 4.2: The average selection acting on each of the seven regions of the HIV2 genome, measured by weighing each expected mutation by its appropriate selection coefficient. Selection on double coding regions appears to tend to be more lenient.

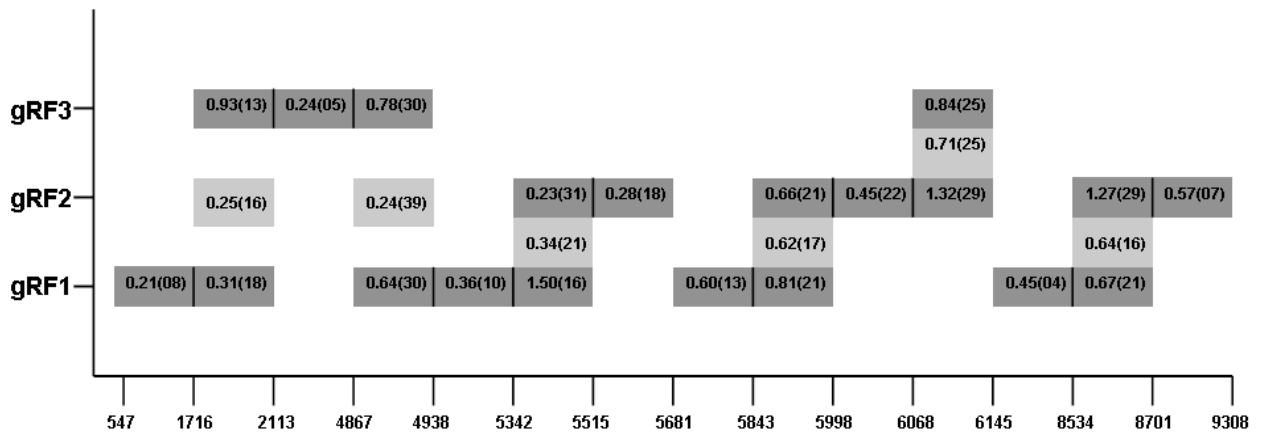


Figure 4.7: The estimated parameters for the seven different genomic regions of the HIV2 U27200 genome based on the HIV2 genomes J04542, M30502, DQ307022 and M15390. Here the darker shades refer to the selection acting on one gene only, and the lighter shades are the selection factors for non-synonymous substitutions in both genes. The error in the parameter estimates (one standard deviation) is given in brackets, in units of  $10^{-2}$ . The x-axis delineates the position on the genome (not to scale).

# Chapter 5

## Discussion & Future Work

*In the course of this thesis we have addressed a number of different questions related to the analysis of viral genomes. We will revise our three main chapters separately and discuss our results, as well as consider drawbacks to our methods and contemplate possible extensions for future research.*

### 5.1 Annotation of Viruses With Non-Conserved Gene Structure

#### 5.1.1 Overview

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 & Jensen, 2005, Meyer & Durbin, 2002]. Albeit just using

evolutionary information and disregarding actual sequence composition — such as codon usage and GC 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 structurally 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 can thus compete favourably with other state of the art methods, as shown in Table 2.8.

Our quality of prediction is highly dependent on sequences not being evolutionarily too close together, in which case our programme 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.

One drawback is our modelling of selection factors. Our assumption of independence between reading frames is one aspect, a problem we address in Chapter 4. An arguably more problematic one, however, is our premise of “one reading frame – one selection factor”. This is not a biologically jus-

tifiable concept, and brings problems along with it, especially when looking at more diverse sequences. Ideally, we would like every genomic region to have its own selection factor. However this would, for  $n$  different regions, result in a  $n^3$  fold increase in state space and is thus, again, not a practical option for the model we are dealing with here. 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 Zanotto *et al.* [1999], de Oliveira *et al.* [2004], Yang & 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 overpowered 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, similar to the method introduced in Chapter 3 — 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 would like to, in the light of Chapter 4, concentrate 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.

We could also incorporate the HKY model instead of the Kimura model into the model, and thus individually capture the mutation of each nucleotide to another as well the effect of base frequency. However, on each sequence we applied our method to, the nucleotide frequency distribution was essentially at equilibrium. Additionally our degeneracy-type-annotation is designed for the use with transitions and transversions. If one marks each nucleotide position as synonymous or non-synonymous for each possible nucleotide mutation, then we may no longer exponentiate the rate matrices, and the subsequent dramatic increase in runtime (see section 5.2) this would bring with it might not make it a suitable payoff.

Including codon bias would be another natural extension to our model and could be accommodated by applying a method similar to the one described in McCauley & Hein [2006]. This would basically involve turning the ancestral sequence into a second order Markov chain, where for each state, given the prior two nucleotides, 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 ORF scanner.

### 5.1.2 Introns

Since our programme can not deal with the presence of introns, there is an inherent degree of failure in our method. The majority of viral genes are encoded by only one exon. There are, however, several viral genes which are split into exons and introns, such as *tat* and *rev* in HIV for example. Note that here, for lack of better terminology, we refer to an intron as something that gets spliced out of a gene. In the case of HIV it is a result of alternative splicing, since the 'introns' actually code for the *env* gene in a different reading frame.

We require the presence of a start and a stop codon to respectively begin and end a coding state, so we are not capable of modelling the latter part of intronic genes with our above method. Instead, we are more likely to either not annotate the region at all as coding, or only pick up on the first exon and extend to annotating it into the intronic region until we hit a 'false' stop codon. However, we may extend our above framework to accommodate for the presence of intronic genes.

We must remember, that a gene may enter an intron in any reading frame and theoretically leave it again in any reading frame. Since the intron gets spliced out, there is generally no need for it to be of length modulo 3. However, in the virus scenario, we may observe that the striving for compactness would suggest the sequence encoding an intron also to be coding for an exon of another gene in a different reading frame (indeed as far as HIV goes, this appears to be the case). This would imply that, were the intron *not* to be of length  $\text{mod } 3$ , we could be merging the two sequences

together once we leave the intron. Let us therefore for theoretical purposes assume that introns in RNA-viruses are always of length  $\text{mod } 3$ , even though further investigation into the biology of alternative splicing in viruses would be necessary to fully justify our model.

Another important point, is that two genes may share an intron with identical splice sites. Indeed, this is more often than not the case in HIV, with the sequence coding for introns in two sequences and an exon in the third. We may again, due to the compactness argument, assume that not all three reading frames switch to the intronic state at one time, and thus introduce merely two transition probabilities belonging to the one and the two sequence switch for both starting and ending an intron.

Let us initially just look at the transitions one can make in a single sequence. Let  $s = (s_1, s_2, s_3)$  be the state vector belonging to one sequence, i.e. representing the vertices on one of the cubes. Let  $e_k$  be a base vector and  $Y_k = e_k \cdot Y$ , with  $(k = 1, 2, 3)$ . Here, for example,  $Y_1 = (Y, 0, 0)$  would relate to an intron being present in the first reading frame and the other two being non-coding. On the other hand  $Y_2 \cdot (e_1 + e_3) = (1, Y, 1)$  would define a coding region in the first and third reading frame, and an intron in the second.

Now let us introduce additional transition probabilities for entering and leaving an intron.

- $\epsilon_{START}/\mu_{START}$  for opening an intron in one/two reading frames.
- $\epsilon_{STOP}/\mu_{STOP}$  for ending an intron in one/two frames.

Then our transition probabilities for entering and leaving an intron are given by the following

**Intron START One**  $s \rightarrow Y_t \cdot e_k$  with  $p = \epsilon_{START}$

if  $s \cdot e_k = 1$  & if  $i \bmod 3 = t$

**Intron START Two**  $s \rightarrow Y_t \cdot (e_k + e_l)$  with  $p = \mu_{START}$

if  $s \cdot e_k = 1$  & if  $s \cdot e_l = 1$  & if  $i \bmod 3 = t$

**Intron STOP One**  $s \rightarrow e_k$  with  $p = \epsilon_{STOP}$

if  $s \cdot e_k = Y_t$  & if  $i \bmod 3 = t$

**Intron STOP Two**  $s \rightarrow e_k + e_l$  with  $p = \mu_{STOP}$

if  $s \cdot e_k = Y_t$  & if  $s \cdot e_l = Y_t$  & if  $i \bmod 3 = t$

Although we are dealing with two sequences, and could theoretically split  $\epsilon_{START}$  into three separate probabilities equivalent to  $\alpha$ ,  $\beta$  and  $\gamma$ , it would make little sense this time. If we are coding in both sequences, then we may switch into an intron in either sequence without any signal, i.e. without anything equivalent to a start codon. We will thus switch into an intronic state if we detect lack of conservation, similarly from intronic to coding when we detect conservation. This will apply to both sequences simultaneously, since transitions are unconditional on the actual sequence and for this reason the above transition probabilities suffice. We may however allow our probabilities  $\epsilon_{START}, \mu_{START}, \epsilon_{STOP}, \mu_{STOP}$  to be drawn from a distribution which accounts for splicing signals. That is to say, if the sequence relates to a wellknown splicing signal, the probability of entering an intron would be modelled to be higher.

Since in each of the three global reading frames we may therefore be in either the non-coding, coding, Intron 1 ( $Y_1$ ), Intron 2 ( $Y_2$ ) or Intron 3 ( $Y_3$ )

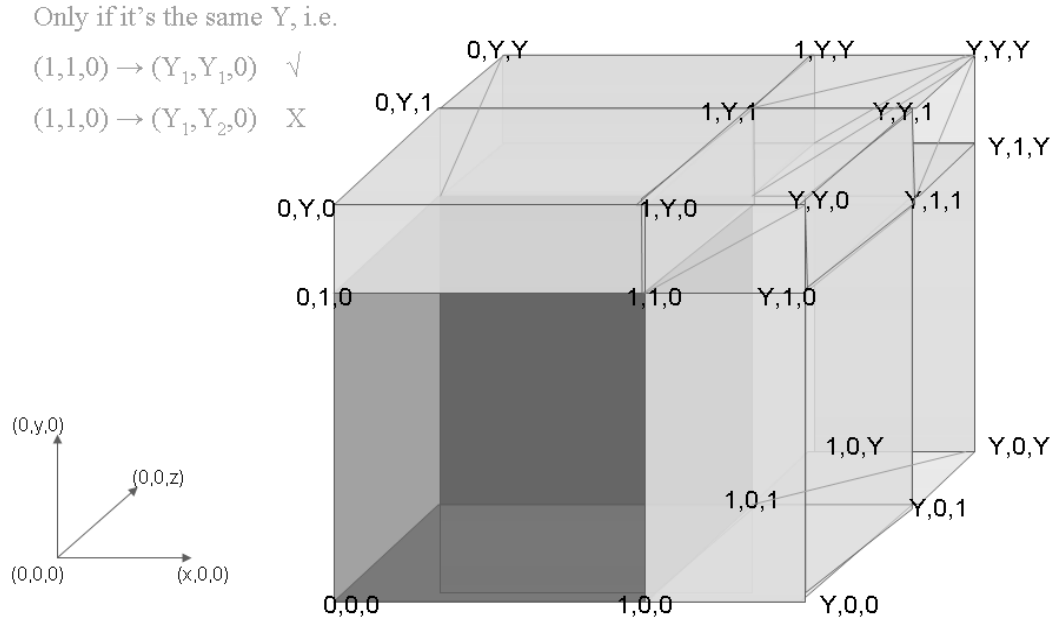


Figure 5.1: Our original cube for one sequence extended to accommodate for introns. We have  $Y = Y_t$  as the intron state for each reading frame, where  $Y_t$  implies having entered at sequence position  $x_i$ , where  $i = t \bmod 3$ . We have also added six transition probabilities relating to the simultaneous opening of two introns in two different reading frames.

state, and due to our not insisting on common gene structure, this model requires  $5^3 \cdot 5^3 = 15,625$  states, which would cause severe complexity issues. It is however debatable, whether the signal provided would be strong enough to detect introns.

### 5.1.3 Ribosomal Slippage

A further important feature is our not drawing ribosomal slippage into account. Ribosomal slippage is a biological phenomenon which sometimes occurs in overlapping genes. The most prominent example is the *gag-pol* polyprotein in HIV. Whilst coding for the *gag* gene the ribosome slips on the sequence strand, resulting in a  $-1$  frameshift and thus the synthesis of

the *gag* and *pol* genes. Ribosomal frameshifting is a very controlled biological event, requiring both a heptameric X XXY YYZ consensus slippery sequence (in HIV this is given by T TTT TTA) and a downstream secondary RNA structure which causes the ribosome to pause (a stem-loop region in HIV). This phenomenon, however, should under optimal conditions be a rare event, occurring only 1 in 10 to 20 times the RNA is translated. The controlled frequency ensures that the synthesis of *gag* and *gag-pol* occurs in the correct ratio required for optimal enzyme activation and virus assembly. To model this feature, we could introduce three additional transition probabilities  $\alpha_{rib}, \beta_{rib}, \gamma_{rib}$  which would be applied upon scanning a sequence of composition X XXY YYZ, as opposed to the usual ATG, in at least one of the aligned genomes.

#### 5.1.4 Multiple Sequences

We naturally would also like to extend our method to multiple sequences, since especially with viral genomes such data is readily at hand. This is however a far from trivial problem, both from a computational and from a model design point of view. We can easily add additional cubes to the state space, but how to define transition probabilities equivalent to our original  $\alpha, \beta$  and  $\gamma$  is less clear. To illustrate this, let us assume we have three sequences, each of which may be in the non-coding (*n*) or coding (*c*) state. The different scenarios we have to consider within a certain reading frame are shown in table 5.1.4.

We therefore need, for the three sequence scenario, to introduce 6 param-

$nnn \rightarrow cnn, ncn, nnc$	$\Leftarrow \alpha_1$
$\rightarrow ccn, cnc, ncc$	$\Leftarrow \alpha_2$
$\rightarrow ccc$	$\Leftarrow \alpha_3$
$cnn \rightarrow ccn$	$\Leftarrow \alpha_4$
$\rightarrow ccc$	$\Leftarrow \alpha_5$
$ncn \rightarrow ncc, ccn$	$\Leftarrow \alpha_4$
$\rightarrow ccc$	$\Leftarrow \alpha_5$
$nnc \rightarrow ncc, cnc$	$\Leftarrow \alpha_4$
$\rightarrow ccc$	$\Leftarrow \alpha_5$
$ccn \rightarrow ccc$	$\Leftarrow \alpha_6$
$cnc \rightarrow ccc$	$\Leftarrow \alpha_6$
$ncc \rightarrow ccc$	$\Leftarrow \alpha_6$

Table 5.1: The different possibilities of transitioning between states for the three sequence scenario. Here  $n$  and  $c$  refer to non-coding and coding respectively.

eters for the different transitions that may now occur — with  $k$  sequences we would need  $\sum_{x=1}^k x$ . Considering the state space for  $n$  sequences will be of size  $8^n$ , this problem will bring complexity issues with it very quickly. On the upside, the extension to a vast amount of multiple sequences is not such an imminent issue for our method, since it is designed to compare genomes far apart. In the case of HIV, for example, only HIV1, HIV2, SIV1 and SIV2 can be seen as structurally significantly different, and thus an extension of our method to four sequences would suffice and the higher order cases need generally not be considered.

## 5.2 Annotation of Selection Strengths in Viruses

### 5.2.1 Overview

We have devised an original method for the comparative annotation of viruses containing overlapping reading frames. Most importantly it accounts for intra- and intergenic differences in selection strength along the genome, thus additionally providing a selection annotation of the virus. Having a multiple alignment as an input it attempts to draw information from the vast amount of sequence data available. We do not rely on a prior GenBank annotation, thus making our method readily applicable to novel virus strands. However, in the case of the gene annotation already being known, we may easily fix it to obtain solely the selection annotation.

Up till now, methods for selection analysis have been restricted to simple studies based on the concept of the  $K_a/K_s$  ratio, which in the presence of overlapping reading frames however loses its power. Mizokami *et al.* [1997], Guyader & Ducray [2002], Hughes & Hughes [2005], Pavesi [2006] and Osiowy *et al.* [2006] use it nonetheless, in an attempt to understand the evolutionary pressures underlying multiple coding regions. Their results are subsequently difficult to interpret, since synonymous substitutions in one reading frame may indeed be non-synonymous and thus restricted in another. This may lead to an underestimation of the ‘true’ synonymous substitution rate and the subsequent false inference of positive selection. The fact that we have incorporated varying selection into a model which accounts for overlapping reading frames is a step towards being able to make truly statistically significant statements about the nature of selection on multiple coding regions.

Of course, it is hard for us to make claims on the accuracy of our method, since no 'true' selection annotation exists. Our simulation results however do show that we can successfully infer parameters for sequences evolved under the same model.

We demonstrate our method on an alignment of 4 HIV2 strains. We discover strong purifying selection acting on non-synonymous substitutions of stop codons, as well as on the majority of the *gag* and *pol*. Several sites within the HIV2 virus appear to be under less stringent negative selection, one in particular corresponding to a known hypervariable region in the *env* gene.

Interestingly enough, in HIV2 the overlapping regions appear to be under less stringent selection, though due to the small sample space it is difficult to draw significant conclusions from this. When analysing the Hepatitis B virus, we can draw bolder conclusions about the evolutionary behaviour of multiple coding regions, due the large amount of coding sites in overlapping reading frames. To the best of our knowledge we are the first to provide a full selection annotation of the Hepatitis B virus genome. We note that here selection in overlapping regions is more conserved. We also investigate differences in the evolutionary pressures underlying the coding regions of Hepatitis B and highlight several regions which appear to be under increased mutational pressure, notably in the S and C genes.

There are several drawbacks to our model. Firstly, the fact that we are fixing nucleotide context over time brings severe problems with it, that are not easily rectified. As mentioned before, Pedersen & Jensen [2001] develop an accurate evolutionary model but need to resort to complicated MCMC

methods in order to estimate parameters. Since complexity is already an issue with our method, fixing the nucleotide context over time is currently our only viable option. Having to exponentiate a large number of matrices results in our running into CPU usage problems and having long runtimes. We could either simplify our model to attain closed form solutions to the matrix exponents, or investigate Taylor expansions in order to find approximate closed form solutions. Secondly, we only account for conservation due to protein coding regions, and thus the evolutionary behaviour of transcription factor binding sites or non-coding RNA genes is currently ignored in our method, and something that future work should consider.

### **5.2.2 RNA secondary structure**

Letting selection vary along the genome is an improvement upon most evolutionary models, but we still neglect the effect of RNA secondary structure on sequence evolution. If we were to incorporate a stochastic context free grammar, as presented in Knudsen & Hein [1999] and discussed in Eddy [2001], into our HMM methodology we could model RNA structure and sequence evolution simultaneously. This would allow us for example to have different evolutionary models for stem and loop regions integrated into our method. Additionally we would be able to model inter-site dependence between different loci, as would be the case for two nucleotides forming a base-pair in a RNA structure stem region.

### 5.2.3 Recombination

Recombination is a complicating factor which is unwisely neglected in fast evolving viral genomes. As shown in Scheffler *et al.* [2006], not accounting for recombination can have a dramatic effect on estimated selection strengths. An obvious extension to our method would be to take an approach similar to Scheffler *et al.* [2006] and computationally detect recombination breakpoints, then let our phylogenetic tree vary between these points. This would allow us to obtain different tree topologies for different parts of the genome. An alternative approach would be to include tree topologies as hidden states in our HMM, and take an approach similar to Husmeier *et al.* [2003] or Minin *et al.* [2005]. One could imagine, given a certain seed tree topology relating to the start state, at each nucleotide position to be able to transition to another tree state with certain restrictions on the new state. For example we could specify a tree topology to be ‘adjacent’ in tree state space if one topology may be obtained from the other by the relocation of one tree branch.

### 5.2.4 Simultaneous Inference of Alignment

A more serious problem is that neither CLUSTALW nor many other alignment programmes take non-coding and coding, yet alone overlapping reading frames, into account when constructing the alignment. Additionally, in the light of the point above, CLUSTALW uses a specific self-inferred tree for the construction of the alignment. How much this influences our results is difficult to say, but ideally we would incorporate alignment and simultaneous tree-building into our method. This is a much more complicated procedure

than the work done in Chapter 4, since here we are dealing with a hidden Markov model. We would yet again have to extend our state space by, in the example of two sequences, splitting each of the existent states into Match-Match, Delete and Insert states and thus align and estimate parameters in one go. For multiple sequences this would obviously result in an even larger state space, and it is likely that we would quickly run into CPU usage and runtime problems.

## **5.3 Investigating Selection: A Statistical Alignment Approach**

### **5.3.1 Overview**

We have introduced a novel method for the estimation of selection strengths not biased by the use of a fixed alignment. By integrating a statistical alignment procedure into our parameter estimation, we are no longer reliant on a fixed alignment input. Instead of working with the observed observations in a set alignment, we rather calculate the expected number of observations, and are thus essentially weighting our parameter estimates by the probability of each possible alignment.

We test our method in a number of different simulation studies against the use of a fixed alignment, which we obtain using CLUSTALW. We show that on average our statistical approach has up to 30% higher absolute sensitivity, and that both evolutionary distance and the length of a double coding region have a lesser effect on our results than when using a fixed alignment.

Our study focuses on trying to understand the selection mechanism underlying overlapping reading frames. With reference to the Hepatitis B genome, which boasts over 1500 multiple coding sites, we address several questions such as the selection a mutation is under, when it causes a non-synonymous mutation in two genes simultaneously. That is to say, if gene  $A$  and gene  $B$  are under selection  $f_A$  and  $f_B$  respectively, will a mutation affecting both necessarily be under selection  $f_A \cdot f_B$ ? A likelihood ratio test between the restricted multiplicative and the full model suggests this is not the case.

We also investigate the strength of selection on double coding regions, with different genomes indicating different results. In Hepatitis B we notice selection on double coding regions not being significantly different to that acting on single coding regions. In HIV2 however, surprisingly, single coding regions appear to be on average under up to two-thirds as stringent selection as double coding regions — a result contrary to the views of Spiropoulou & Nichol [1993], Walewski *et al.* [2001], but supported by the recent papers published by McCauley *et al.* [2007] and de Oliveira *et al.* [2004].

Another feature which is particular to our method, is that we may separate selection acting on the different reading frames in an overlapping region. We find especially in HIV2 a certain division of selection occurring, similar to that observed in Potato Leafroll Virus by Guyader & Ducray [2002] and in Microviridae by Pavesi [2006]. Essentially, it appears as though in an overlapping region one gene can take over the fast evolving function, whilst the other's behaviour is more conserved. Since this is not something we observed in our simulation studies, it does not seem to be an artefact of our model.

It is however not a counterintuitive phenomenon — when an overlapping

region is ‘created’ by the elongation of one of the genes involved or otherwise, then it is likely to be initially under non-negative selection. Since the organism survived both with and without the overlap, it should have the freedom to evolve without detrimental effects. A thus logical natural process would be for the newly coding region to evolve quickly, testing out various evolutionary niches that may be beneficial, whilst the other gene remains under similar selection as before. The estimated selection strengths may subsequently help deduce which overlaps are the ‘newer’ regions — for example our study suggests that the *pol* gene extended itself both onto the *pol* and the *vif* gene. The effect would essentially be similar to that noted on selection occurring on duplicated genes, where the duplicated gene reaches fixation in the population due to initially being under positive selection [Zhang, 2004].

Up till now, as mentioned above, other methods dealing with related issues have made use of the concept of  $K_a/K_s$  ratio, which however creates problems when applied to overlapping reading frames. For this reason we decided on the HMM based approach presented in chapter 3 and estimated selection as acting on a single nucleotide basis, but at the cost of not being able to pry apart selection acting on different reading frames. Most importantly however, all of the above methods use a fixed alignment and are thus prone to a great variability in their estimated parameters, dependent on the alignment. Our method alone manages to circumvent this problem by using a statistical approach, and thus we remove any uncertainty caused by the use of one particular alignment simply by considering all alignments. The improvement we observe by doing this, encourages us even to suggest that the use of fixed alignments generally should be avoided where possible, and

statistical alignments become not the exception but the norm.

### 5.3.2 Varying Transition and Transversion Rates

One drawback to our method is the fact that for each descendent sequence we model transition and transversion rates as constant along the genome. Ideally, we would let these change as well, in order to account for constraining elements such as RNA secondary structure. For example, we could imagine parts of the genome, belonging to stem regions in the RNA secondary structure, evolving slower than loop regions. Mirroring our handling of different selection strengths in chapter 3, we could introduce a set of  $v$  evolutionary speeds for each position to choose from, together with an autocorrelation factor  $\nu$ , and thus account for heterogenic evolutionary rates. We would be incorporating additional states into our current methodology, where in our alignment procedure the indel model would have  $v$  orthogonal layers corresponding to our  $v$  evolutionary speeds. We would have a transition probability of  $\nu$  from each state to itself and of  $\frac{1-\nu}{v-1}$  to one of the others. We would thus obtain expected counts which accommodated for intragenomic difference in transition and transversion rates and could proceed in our selection strength estimation as usual.

### 5.3.3 Multiple Alignment

Currently we perform a multiple pairwise alignment using a star shaped tree to compose the joint likelihood function used in our parameter estimation step. This however is clearly a gross simplification of the circumstances,

and ideally we would incorporate a multiple alignment procedure into our method. If we were to fix the phylogenetic tree underlying the sequences, we would simply have to extend our alignment model to one which deals with multiple sequences and could construct our likelihood function as before. A further extension would be to incorporate a tree-estimation procedure into our methodology, and thus not be reliant on a fixed phylogeny.

### 5.3.4 Organic Choosing of Breakpoints

A major problem worth considering is our fixing a partition prior to analysis. It would be even more interesting to be able to incorporate some sort of hidden procedure, in which breakpoints between regions were chosen organically from the data. We could then truly start questioning which parts of the genome behaved in different ways, as opposed to being restricted to the ‘trial and error’ approach that is the essence of our method now. One way to approach the problem is to *a priori* set a maximal number of breakpoints,  $|P|$  say. We would incorporate a chain of  $|P| + 1$  states belonging to each of the regions into our indel model, similar to the scenario in section 5.3.2. Transitions from one region to itself would occur with probability  $\phi$  and to the following region at any point in the sequence with transition rate  $1 - \phi$ . We would thus in our alignment procedure also be summing over all possible partition annotations and finally use the Viterbi Algorithm to annotate the genome and infer the respective selection strengths for each region.

### 5.3.5 Summary

We have investigated a number of different questions related to the analysis of viral genomes. A model which accounts for overlapping reading frames and the change in gene structure over time has been developed and incorporated into a pairwise *ab initio* annotation procedure for viral genomes. We have also provided a method for the selectional annotation of a viral genome on a nucleotide basis and analysed HIV2 and Hepatitis B genomes with it. Finally, we have devised a statistical alignment method for viral genomes, which for an *a priori* partitioning of the genome infers selection factors for the designated regions without the use of a fixed alignment. Moreover we discover the overlapping regions in HIV2 to be under less stringent selection than the single coding ones, as well as a tendency for a slow evolving region overlapping a fast evolving one. We therefore hope that this thesis has made a valid academic contribution and filled some of the gaps present in our understanding of the evolution of viral genomes.

# Appendix A

## Algorithms

We will introduce three dynamic programming algorithms which as an input have an observed sequence of length  $K$  and an HMM with  $L$  states. Assume our sequence and HMM is as follows

- The observed sequence is  $x_{(1)}, x_{(2)}, \dots, x_{(K)}$ .
- Each of the  $L$  states emits  $M$  symbols with probability  $e_{ml}$  ( $1 < l < L; 1 < m < M$ ).
- We have a transition matrix  $a_{ij}$  for switching from state  $i$  to state  $j$ .
- Let  $\pi = (\pi_k)$  be the true state path through the sequence

### A.1 The Viterbi Algorithm

Given an HMM we are interested in finding which path through all the possible states is the most likely. For this we use the Viterbi algorithm, which relies on the concept of dynamic programming, which in  $O(NL)$  time calculates the most likely state path through the emitted sequence using the following recursions

**Initialization**  $Viterbi_0(0) = 1; Viterbi_0(i) = 0$  ( $1 < i < L$ )

**Recursion**  $Viterbi_{k+1}(j) = \max_i Viterbi_k(i) \cdot a_{ij} \cdot e_j^{x_{(k+1)}}$

**Pointer**  $Pointer_{k+1}(j) = \operatorname{argmax}_i Viterbi_k(i) \cdot a_{ij}$

**Termination**  $StateAnno(K) = \operatorname{argmax}_i Viterbi_K(i)$

**State Annotation**  $StateAnno(k) = Pointer_{k+1}(StateAnno(k+1))$

$Viterbi_k(j)$  thus represents the probability of the most likely path up to the  $k^{th}$  locus ending in state  $j$ . We then for the  $(k + 1)^{th}$  locus obtain  $Viterbi_{k+1}(j)$  and  $Pointer_{k+1}(j)$  by

- for all  $i$  calculating the transition probability  $a_{ij}$  from state  $i$  to  $j$ ,
- multiplying this by  $Viterbi_k(i)$ ,
- multiplying this by the probability of state  $j$  emitting nucleotide  $x_{k+1}$ ,
- taking the maximum over all  $i$ , and finally
- letting this product equal  $Viterbi_{k+1}(j)$  and the maximizing  $i$  equal  $Pointer_{k+1}(j)$ .

This is a dynamic programming algorithm, for every optimal subpath is in itself optimal. Imagine we have an optimal path  $S$  through the sequence with a subpath  $s_1$  ending in state  $i$ . If  $s_1$  weren't itself optimal we could find another subpath  $s_2$  and replace it in  $S$  to achieve an overall even better path. Thus we find the most likely state to have emitted the last locus  $x_K$  and then use the pointers to answer the question "If I am here, what is the most likely path by which I arrived?" Hereby we find the most likely path by travelling backwards through the dynamic programming matrix we have built, using the pointers to guide our way.

## A.2 The Forward Algorithm

The Forward algorithm calculates  $f_i(k)$ , the sum of the probabilities of all paths ending in the observed emission at sequence position  $k$  in state  $i$  using the following recursions:

**Initialization**  $f_0(0) = 0$ ;  $f_i(0) = 1$  ( $i = 1 \dots L$ )

**Recursion**  $f_j(k) = e^{x^{(k)}j} \cdot \sum_i f_i(k - 1) \cdot a_{ij}$  ( $j = 0 \dots L$ ;  $k = 1 \dots K$ );

**Termination**  $P(x) = \sum_i f_i(K) \cdot a_{i0}$

$P(x)$  then represents the probability of observing the entire sequence under the model.

### A.3 The Backward Algorithm

The Backward algorithm calculates  $b_i(k)$ , the sum of the probabilities of all paths beginning in the observed emission at sequence position  $k$  in state  $i$  using the following recursions:

**Initialization**  $b_i(K) = a_{i0}$  ( $i = 0 \dots L$ );

**Recursion**  $b_i(k) = \sum_j a_{ij} \cdot e_j(k+1) \cdot b_j(k+1)$  ( $i = 0 \dots L; k = K-1 \dots 1$ )

**Termination**  $P(x) = \sum_j a_{0j} \cdot e_j(1) \cdot b_j(1)$

As with the forward algorithm,  $P(x)$  represents the probability of observing the entire sequence under the model. We may then use both the above to consider the true path  $\pi$  and calculate the posterior decoding probability of locus  $k$  having been encoded in state  $i$ :

$$P(\pi_k = i|x) = \frac{f_i(k) \cdot b_i(k)}{P(x)} \quad (\text{A.1})$$

### A.4 The Baum Welch Algorithm

Using the Viterbi Algorithm A.1, we may now find the best state path  $\pi^*$  through the sequence — but only for a given set of parameters. Ideally one would like to have just the observed sequence  $x_{(1)}, x_{(2)}, \dots, x_{(K)}$  as an input, and in an iterative procedure deduce both the parameters and the best path from there. The Baum-Welch algorithm does exactly that. We start off with a set of seed parameters  $a_{ij}$  and  $e_i(k)$  and calculate the Forward A.2, Backward A.3 and Sequence probabilities A.1 — then the probability that the transition  $a_{ij}$  was used at locus  $k$  is given by

$$P(p_k = i, p_{k+1} = j|x) = \frac{f_i(k) \cdot a_{ij} \cdot e_j(k+1) \cdot b_j(k+1)}{P(x)} \quad (\text{A.2})$$

Thus the expected number of times that  $a_{ij}$  was used throughout the entire sequence is

$$A_{ij} = \sum_k \frac{f_i(k) \cdot a_{ij} \cdot e_j(k+1) \cdot b_j(k+1)}{P(x)} \quad (\text{A.3})$$

If all transition probabilities are free parameters, we may deduce from this that the maximum likelihood estimators for the transition probabilities  $a_{ij}$  are given by

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{\tilde{j}} A_{i\tilde{j}}} \quad (\text{A.4})$$

Similarly we may calculate the new estimated emission parameters, and recalculate the new forward, backward and sequence probabilities. Once the sequence probability  $P(x)$  has not changed more than a certain threshold after iteration, we stop the procedure and use the Viterbi algorithm with the final parameters as an input to obtain the most likely path through the observed sequence.

## A.5 The Newton Raphson Iteration

Suppose we are given a one-dimensional function  $f(x)$  whose local maximum  $x^*$  we are supposed to find. Since stationary points of  $f$  will be roots of its derivative  $f'$  we may instead attempt to solve for these. Thus if we start off with an initial guess  $x_0$  and we assume that  $f$  is twice differentiable at this point, then using Taylor's Theorem we may write for the expansion around the point  $x_0$

$$f'(x_0 + d) = f'(x_0) + f''(x_0) \cdot d + \frac{f'''(x_0)}{2} \cdot d^2 + \dots \quad (\text{A.5})$$

For small values of  $d$  and well behaved functions we may deduce that if  $f'(x_0 + d) = 0$

$$d = -\frac{f'(x_0)}{f''(x_0)} \quad (\text{A.6})$$

So updating  $x_0$  by  $x_0 - d$  we are thus edging ourselves closer towards the local maximum, our iteration thus being given by

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \quad (\text{A.7})$$

Clearly we are not guaranteed to find the global maximum, and which point we reach after several iterations greatly depends on our initial starting value  $x_0$ . However with an adequate guess at a starting point close enough to  $x^*$  and a not too "hilly" function we are guaranteed to converge to the global maximum  $x^*$ .

The above iterative scheme easily generalizes to the multiple dimensional case where we are dealing with a function  $f(x_1, \dots, x_n)$  of several variables which we will denote by  $f(x)$ . Replacing the derivative  $f'$  with the gradient  $\nabla f(x)$  and the second derivative  $f''$  with the Hessian matrix  $\mathbf{H}f(x)$  we may reformulate the iteration as

$$x_{n+1} = x_n - [\mathbf{H}f(x_n)]^{-1} \nabla f(x_n) \quad (\text{A.8})$$

where  $x_0$  is our initial seed vector.

# Bibliography

- Alexandersson,M., Cawley,S., Pachter,L. (2003) SLAM: Cross-Species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model, *Genome Research*, **13(3)**, 496-502.
- Barrell III,B.G., Air,G.M., Hutchison,C.A. (1976) Overlapping genes in bacteriophage  $\phi$ X174 *Nature*, **264**, 3441.
- Batshake, B., Sundelin, J., (1996) The mouse genes for the EP1 prostanoid receptor and the PKN protein kinase overlap, *Biochemical and Biophysical Research Communications*, **227**, 7076.
- Beck,D.L., Guildford,P.J., Voot,D.M., Anderson,M.T., Forest,R.L. (1991) Triple gene block proteins of white clover mosaic potexvirus are required for transport, *Virology*, **183**, 695-702.
- Belshaw,R., Pybus,O.G., Rambaut,A. (2007) The evolution of genome compression an dgenomic novelty in RNA viruses, *Genome Research*, **In Press**
- Besemer,J., Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding, *Nucleic Acids Research*,**27(19)**, 3911-3920.
- Besemer,J., Lomsadze,A., Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, *Nucleic Acids Research*, **29(12)**, 2607-2618.
- Bristow,J., Tee,M.K., Gitelman,S.E., Mellon,S.H., Miller,W.L. (1993) Tenascin-X: a novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B, *Journal of Cellular Biology*, **122**, 265278.
- Burch,C.L., Turner,P.E., Hanley,K.A. (2003) Patterns of epistasis in RNA viruses: a review of the evidence from vaccine design, *Journal of Evolutionary Biology*, **16**, 1223-1235.

- Burge,C., Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA, *Journal of Molecular Biology*, **268**, 78-94
- Cann,A.J (1997) Principles of Molecular Virology, *Academic Press*, **2nd Ed**, Chapter 3.
- Chain,B.M., Myers,R. (2005) Variability and conservation in hepatitis B virus core protein, *BMC Microbiology*, **5(33)**.
- Coffin,J., Hughes,S., Varmus,H. (1997) "Retroviruses" *Cold Spring Harbor Laboratory Press*
- Cooper,P.R., Smilinich,N.J., Day,C.D., Nowak,N.J., Reid,L.H., Pearsall,R.S., Reece,M., Prawitt,D., Landers,J., Housman,D.E., Winterpacht,A., Zabel,B.U., Pelletier,J., Weissman,B.E., Shows,T.B., Higgins,M.J. (1998) Divergently transcribed overlapping genes expressed in liver and kidney and located in the 11p15.5 imprinted domain, *Genomics* **49**, 3851.
- Crotty,S., Cameron,C.E., Andino,R. (2001) RNA virus error catastrophe: Direct molecular test by using ribavirin, *PNAS USA* **98**, 6895-6900.
- de Groot,S., Mailund,T., Hein,J. (2007) Comparative Annotation of Viral Genomes with Non-Conserved Gene Structure, *Bioinformatics*, **23(9)**, 1080-1089.
- de Groot,S., Mailund,T., Lunter,G.A., Hein,J. (2007) Investigating Selection in Viruses: A Statistical Alignment Approach, *Bioinformatics*, In Review.
- de Oliveira,T., Salemi,M., Gordon,M., Vandamme,A-M., van Rensburg,E.J., Engelbrecht,S., Coovadia,H.M., Cassol,S. (2004) Mapping Sites of Positive Selection and Amino Acid Diversification in the HIV Genome, *Genetics*, **167**, 1047-1058.
- de Zanolto,P., Kallas,E., de Souza,R., Holmes,E. (1999) Genealogical Evidence for Positive Selection in the nef Gene of HIV-1, *Genetics*, **153**, 1077-1089.
- Ding,S.W., Anderson,B.J., Haase,H.R., Symons,R.H. (1994) New overlapping gene encoded by the cucumber mosaic virus genome, *Virology*, **198(2)**, 593-601.
- Drake,J.W. and Holland,J.J. (1999) Mutation rates among RNA viruses, *PNA USA*, **96**, 13910-13913.

- Duhig,T., Ruhrberg,C., Mor,O., Fried,M. (1998), The human Surfeit locus, *Genomics*, **52**, 7278.
- Durbin,R., Eddy,S., Krogh,A., Mitchison,G. (1998) Biological Sequence Analysis, *Cambridge University Press*.
- Eddy,S.R. (2001) How do RNA-folding algorithms work? *Nature Biotechnology*, **22**, 1457-1463.
- Edgar,A.J. (2003), The gene structure and expression of human ABHD1: overlapping polyadenylation signal sequence with Sec12, *BMC Genomics*, **4**, 18.
- Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**, 113
- Eigen, M. (1971) Selforganization of matter and evolution of biological macromolecules, *Naturwissenschaften*, **58**, 465-523.
- Elena,S.F., Carrasco,P., Daros,J.A., and Sanjuan,R. (2006) Mechanisms of genetic robustness in RNA viruses, *EMBO Reports*, **7**, 168-173.
- Felsenstein,J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2), *Cladistics*, **5**, 164-166.
- Fiddes,A., Godson,A. (1979) Evolution of the three overlapping gene systems in G4 and  $\phi$ X174, *Journal of Molecular Biology*, **133**, 19-43.
- Firth,A.E., Brown,C.M. (2005) Detecting overlapping coding sequences with pairwise alignments, *Bioinformatics*, **21(3)**, 282-292.
- Firth,A.E., Brown,C.M. (2006) Detecting overlapping coding sequences in virus genomes, *BMC Bioinformatics*, **7(75)**.
- Fukuda,Y., Nakayama,Y., Tomita,M. (2003) On dynamics of overlapping genes in bacterial genomes, *Gene*, **323**, 181-187.
- Gamerman,D. (1997) Markov Chain Monte Carlo, *Chapman & Hill*, London
- Giorgi,C., Blumberg,B.M., Kolakofsky,D. (1983) Sendai virus contains overlapping genes expressed from a single mRNA, *Cell* **35:3(2)**, 829-836.
- Goldman,N., Yang,Z.H. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular Biology and Evolution*, **11**, 725-736

- Guyader,S., Ducray,D.G. (2002) Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products, *Journal of General Virology*, **83**, 1799-1807.
- Hasegawa,M., Kishino,H., Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, 160-174
- Hein,J., Støvlbæk,J. (1994) Genomic Alignment, *Journal of Molecular Evolution*, **38**, 310-316.
- Hein,J., Støvlbæk,J. (1995) A maximum-likelihood approach to analyzing non-overlapping and overlapping reading frame, *Journal of Molecular Evolution*, **40**(2), 181-189.
- Hein,J., Støvlbæk,J. (1996) Combined DNA and Protein Alignment, *Methods in Enzymology*, **266**, 402-418.
- Hobolth,A., Jensen,J.L. (2005) Applications of hidden markov models for characterization of homologous DNA sequences with a common gene, *Journal of Computational Biology*, **12**, 186-203.
- Holmes,E.C. (2003) Error thresholds and the constraints to RNA virus evolution, *Trends in Microbiology*, **11**, 543-546.
- Hughes,A.L., Westover,K., Da Silva,J., O'Connor,D.H., Watkins,D.I. (2001) Simultaneous positive and purifying selection on overlapping reading frames of the *tat* and *vpr* genes of simian immunodeficiency virus, *Journal of Virology*, **75**(17), 7966-7972.
- Hughes,A.L., Hughes,M.A.K. (2005) Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes, *Virus Research*, **113**, 81-88.
- Husmeier,D., McGuire,G. (2003) Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo, *Molecular Biological Evolution*, **20**(3), 315-337.
- Jenkins,G.M., Rambaut,A., Pybus,O.G., and Holmes,E.C. (2002) Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis, *Journal of Molecular Evolution*, **54**, 156-165.
- Johnson,Z.I., Chisholm,S. (2006) Properties of overlapping genes are conserved across microbial genomes, *Genome Research*, **14**, 2268-2272.

- Jukes,T.H., Cantor,C.R. (1969) Evolution of Protein Molecules, In: H.N. Munro(editor) *Mammalian Protein Metabolism* Academic Press, New York, 21-132.
- Kasper,G., Taudien,S., Staub,E., Mennerich,D., Rieder,M., Hinzmann,B., Dahl,E., Schwidetzky,U., Rosenthal,A., Rump,A. (2002) Different structural organization of the encephalopsin gene in man and mouse, *Gene*, **295**, 2732.
- Kennerson,M.L., Nassif,N.T., Dawkins,J.L., DeKroon,R.M., Yang,J.G., Nicholson,G.A. (1997) The Charcot-Marie-Tooth binary repeat contains a gene transcribed from the opposite strand of a partially duplicated region of the COX10 gene, *Genomics*, **46**, 6169.
- Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution*, **16**, 111-120.
- Knudsen,B., Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history, *Bioinformatics*, **15**, 446-454.
- Kiyosawa,H., Abe,K. (2002) Speculations on the role of natural antisense transcripts in mammalian X chromosome evolution, *Cytogenetic Genome Research*, **99**, 151-156.
- Korf,I., Flicek,P., Duan,D., and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction, *Bioinformatics*, **17**, S140-S148.
- Kozlov,N.N. (2000) Overlapping Genes and Variability of the Genetic Code, *Doklady Biological Sciences*, **375**, 677-680.
- Kozlov,N.N. (2000) Analysis of a Set of Overlapping Genes, *Doklady Biochemistry*, **373**, 119-122.
- Laabi,Y., Gras,M.P., Brouet,J.C., Berger,R., Larsen,C.J., Tsapis,A. (1994) The BCMA gene, preferentially expressed during B lymphoid maturation, is bidirectionally transcribed, *Nucleic Acids Research*, **22**, 1147-1154.
- Li,W.S., Wu,C.-I., Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes, *Molecular Biology and Evolution* **2**(2), 150-174.

- Li,W.S. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution, *Journal of Molecular Evolution* **36**, 96-99.
- Liu,B., Dou,C.L., Prabhu,L., Lai,E. (1999) FAST-2 is a mammalian winged-helix protein which mediates transforming growth factor beta signals, *Molecular Cellular Biology*, **19**, 424430.
- Lukashin,A., Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding *Nucleic Acids Research*, **26(4)**, 1107-1115.
- Lunter,G.A., Drummond,A.J., Miklós,I., Hein,J. (2004) Statistical Alignment: Recent Progress, New Applications, and Challenges: Statistical methods in Molecular Evolution, *Springer Verlag's Series in Statistics in Health and Medicine*, 2004.
- Lunter,G.A. (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes, *Bioinformatics*, **23(13)**, i289-i296.
- Lunter,G.A. (2007) HMMoC - a compiler for hidden Markov models, *Bioinformatics*, **23(18)**, 2485-2487.
- Majoros W.H., Pertea M., Salzberg S.L. (2005) Efficient implementation of a Generalized Pair Hidden Markov Model for comparative gene finding, *Bioinformatics*, **21**, 1782-1788.
- Makalowska,I., Lin,C.-F., Makalowski,W. (2005) Overlapping genes in vertebrate genomes, *Computational Biology and Chemistry*, **29**, 112.
- Malavasic,M.J., Elder,R.T. (1990) Complementary transcripts from two genes necessary for normal meiosis in the yeast *Saccharomyces cerevisiae*, *Molecular Cellular Biology*, **10**, 28092819.
- Mansky, L.M. (2000) In vivo analysis of human T-cell leukemia virus type 1 reverse transcription accuracy, *Journal of Virology*, **74**, 9525-9531.
- Mayo,M.A., Robinson,D.J., Jolly,C.A., Hyman,L. (1989) Nucleotide sequence of potato leafroll luteovirus RNA, *Journal of General Virology*, **70**, 1037-1051.
- McCauley,S., Hein,J. (2006) Using HMMs and observed evolution to annotate viral genomes, *Bioinformatics*, Advance Access published online on April 13, 2006.
- McCauley,S., de Groot,S., Mailund,T., Hein,J. (2007) Annotation of Selection Strengths in Viral Genomes, *Bioinformatics*, Advance Access.

- Metzler,D., Fleiner,R., Wakolbinger,A., von Haeseler,A. (2001) Assessing Variability by Joint Sampling of Alignments and Mutation Rates, *Journal of Molecular Evolution*, **53(1)**, 660-669.
- Meyer,I.M., Durbin,R. (2002) Comparative ab initio prediction of gene structure using pair HMMs, *Bioinformatics*, **18(10)**, 1309-1318.
- Mills,R., Rozanov,M., Lomsadze,A., Tatusova,T., Borodovsky,M. (2003) Improving gene annotation of complete viral genomes, *Nucleic Acid Research*, **31(23)**, 7041-7055.
- Minin,V.N., Dorman,K.S., Fang,F. and Suchard,M.A. (2005) Dual multiple change-point model leads to more accurate recombination detection *Bioinformatics*, **21**, 3034-3042.
- Misener,S.R., Walker,V.K. (2000) Extraordinarily high density of unrelated genes showing overlapping and intraintronic transcription units, *Biochimica et Biophysica Acta*, **1492**, 269270.
- Misra,S., Crosby,M.A., Mungall,C.J., Matthews,B.B., Campbell,K.S., Hradecky,P., Huang,Y., Kaminker,J.S., Millburn,G.H., Prochnik,S.E., Smith,C.D., Tupy,J.L., Whitfield,E.J., Bayraktaroglu,L., Berman,B.P., Bettencourt,B.R., Celniker,S.E., de Grey,A.D., Drysdale,R.A., Harris,N.L., Richter,J., Russo,S., Schroeder,A.J., Shu,S.Q., Stapleton,M., Yamada,C., Ashburner,M., Gelbart,W.M., Rubin,G.M., Lewis,S.E. (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review, *Genome Biology*, **3**, RESEARCH0083.
- Mizokami,M., Orito,E., Ohba,K., Ikeo,K., Lau,J.Y., Gojobori,T. (1997) Constrained evolution with respect to gene overlap of Hepatitis B Virus, *Journal of Molecular evolution*, **44(1S)**, 83-90.
- Morch,M.D., Boyer,J.C., Haenni,A.L. (1988) Overlapping open reading frames revealed by complete nucleotide sequencing of turnop yellow mosaic virus genomic RNA, *Nucleic Acids Research*, **16**, 6157-6173.
- Morel,Y., Bristow,J., Gitelman,S.E., Miller,W.L. (1989) Transcript encoded on the opposite strand of the human steroid 21-hydroxylase/complement component C4 gene locus, *PNAS U.S.A.*, **86**, 65826586.
- Morrison,D.F. (1976) *Multivariate Statistical Methods*, *New York:McGraw-Hill*.

- Narechania,A., Terai,M., Burk,R.D. (2005) Overlapping reading frames in closely related human papillomaviruses result in modular rates of selection within E2, *Journal of General Virology*, **86**, 1307-1313.
- Nei,M., Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Molecular Biology and Evolution*, **3**, 418-426.
- Nicolaidis,N.C., Kinzler,K.W., Vogelstein,B. (1995) Analysis of the 5' region of PMS2 reveals heterogeneous transcripts and a novel overlapping gene, *Genomics*, **29**, 329334.
- Nowak,M.A. (1992) What is a Quasispecies? *Trends in Ecology and Evolution*, **7**, 118-121.
- Ohinata,Y., Sutou,S., Kondo,M., Takahashi,T., Mitsui,Y. (2002) Male-enhanced antigen-1 gene flanked by two overlapping genes is expressed in late spermatogenesis, *Biology of Reproduction*, **67**, 18241831.
- Osiowy,C., Giles,E., Tanaka,Y., Mizokami,M., Minuk,G.Y. (2006) Molecular Evolution of Hepatitis B Virus over 25 Years, *Journal of Virology*, **80(21)**, 10307-10314.
- Pavesi,A., De Iaco,B., Granero,M.I., Poratei,A. (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses, *Journal of Molecular Evolution*, **44(6)**, 625-631.
- Pavesi,A. (2000) Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus, *Journal of Molecular Evolution*, **50(3)**, 284-295.
- Pavesi,A. (2006) Origin and evolution of overlapping genes in the family *Microviridae*, *Journal of General Virology*, **87**, 1013-1017.
- Pedersen,J.S., Hein,J. (2003) Gene finding with a hidden Markov model of genome structure and evolution, *Bioinformatics*, **19(2)**, 219-227.
- Pedersen,A.M., Jensen,J.L. (2001) A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames, *Molecular Biological Evolution*, **18(5)**, 763-776.
- Peterson,J.A., Myers,A.M. (1993) Functional analysis of mRNA 3' end formation signals in the convergent and overlapping transcription units of

- the *S. cerevisiae* genes RHO1 and MRP2, *Nucleic Acids Research*, **21**, 55005508.
- Petrukhin,K., Koisti,M.J., Bakall,B., Li,W., Xie,G., Marknell,T., Sandgren,O., Forsman,K., Holmgren,G., Andreasson,S., Vujic,M., Bergen,A.A., McGarty-Dugan,V., Figueroa,D., Austin,C.P., Metzker,M.L., Caskey,C.T., Wadelius,C. (1998) Identification of the gene responsible for Best macular dystrophy, *Nature Genetics*, **19**, 241247.
- Rogozin,I., Spiridinov,A.N., Sorokin,A.V., Wolf,Y.I., Jordan,I.K., Tatusov,R.L., Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes, *Trends in Genetics*, **18(5)**, 228-232.
- Salzberg,S.L., Delcher,A.L., Kasif,S., White,O. (1998) Microbial gene identification using interpolated Markov models, *Nucleic Acids Research*, **26(2)**, 544-548.
- Samuel,A. (1989) Polycistronic animal virus mRNAs, *Progress in Nucleic Acids Research and Molecular Biology*, **37**, 127-153.
- Sanjuan,R., Moya,A., Elena,S.F. (2004) The distribution of fitness effects cause by single-nucleotide substitutions in RNA viruses, *PNAS USA*, **102**, 8396-8401.
- Scheffler,K., Martin,D.P., Seoighe,C. (2006) Robust inference of positive selection from recombining coding sequences, *Bioinformatics*, textbf22(20), 2493–2499.
- Seibert,S., Howell,C., Hughes,M., Hughes,A. (1995) Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1), *Molecular Biology and Evolution*, **12(5)**, 803-813.
- Seligmann,H., Pollock,D.D. (2004) The ambush hypothesis: hidden stop codons prevent off frame gene reading, *DNA and Cell Biology*, **23**, 701-705.
- Siepel,A., Haussler,D. (2004) Combining phylogenetic and hidden Markov models in biosequence Analysis, *Journal of Computational Biology*, **11(2-3)**, 413-428.
- Shepherd,J.C.W. (1981) Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code, *Journal Molecular Evolution*, **17**, 94-102.

- Simmonds,P., Balfe,P., Ludlam,C.A., Bishop,J.O., and Brown,A.J. (1990) Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1, *Journal of Virology*, **64**(12), 5840-5850.
- Spencer,C.A., Gietz,R.D., Hodgetts,R.B. (1986) Overlapping transcription units in the dopa decarboxylase region of *Drosophila*, *Nature*, **322**, 279-281.
- Spiropoulou,C.F., Nichol,S.T. (1993) A small highly basic protein is encoded in overlapping reading frame within the P gene of vesicular stomatitis virus, *Journal of Virology*, **67**, 3103-3110.
- Steinhauer,D.A., Domingo,E., and Holland,J.J. (1992) Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase, *Gene*, **122**, 281-288.
- Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) Some mathematical questions in biology DNA sequence analysis, *American Mathematic Society*, Providence, 57-86.
- Thompson,J.D., Higgins,D.G., Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22**(22), 4673-4680.
- Thorne, J.L., Kishino,H., Felsenstein,J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences , *Journal of Molecular Evolution*, **33**, 114-124.
- Tvrđik,P., Asadi,A., Kozak,L.P., Nuglozeh,E., Parente,F., Nedergaard,J., Jacobsson,A. (1999) Cig30 and Pitx3 genes are arranged in a partially overlapping tail-to-tail array resulting in complementary transcripts, *Journal of Biological Chemistry*, **274**, 26387-26392.
- Yang,Z.H., Swanson,W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes, *Molecular Biology and Evolution*, **19**(1), 49-57.
- Yang,Z.H., Nielsen,R., Goldman,N., Krabbe Pedersen,A.-M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites, *Genetics*, **155**, 431-449.

Walewski,J.L., Keller,T.R., Stump,D.D., Branch,A.D. (2001) Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame, *RNA*, **7(5)**, 710-721.

Williams,T., Fried,M. (1986) A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends, *Nature*, **322**, 275-279.

Zhang,J. (2004) The infancy of duplicate genes, *Heredity*, **92**, 479-480.

Zhou, C., Blumberg, B., 2003. Overlapping gene structure of human VLCAD and DLG4. *Gene* 305, 161-166.

Zuker,M. (1991) Suboptimal sequence alignment in molecular biology. Alignment with error analysis, *Journal of Molecular Biology*, **221**, 403-420.

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/>