BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 3 2005, pages 282–292
doi:10.1093/bioinformatics/bti007

Detecting overlapping coding sequences with pairwise alignments

Andrew E. Firth and Chris M. Brown[*]

Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received on May 20, 2004; revised on July 29, 2004; accepted on August 26, 2004
Advance Access publication September 3, 2004

ABSTRACT

Motivation: Overlapping gene coding sequences (CDSs) are particularly common in viruses but also occur in more complex genomes. Detecting such genes with conventional gene-finding algorithms can be difficult for several reasons. If an overlapping CDS is on the same read-strand as a known CDS, then there may not be a distinct promoter or mRNA. Furthermore, the constraints imposed by double-coding can result in atypical codon biases. However, these same constraints lead to particular mutation patterns that may be detectable in sequence alignments.

Results: In this paper, we investigate several statistics for detecting double-coding sequences with pairwise alignments—including a new maximum-likelihood method. We also develop a model for double-coding sequence evolution. Using simulated sequences generated with the model, we characterize the distribution of each statistic as a function of sequence composition, length, divergence time and double-coding frame. Using these results, we develop several algorithms for detecting overlapping CDSs.

The algorithms were tested on known overlapping CDSs and other overlapping open reading frames (ORFs) in the hepatitis B virus (HBV), *Escherichia coli* and *Salmonella typhimurium* genomes. The algorithms should prove useful for detecting novel overlapping genes—especially short coding ORFs in viruses.

Availability: Programs may be obtained from the authors.

Contact: chris.brown@otago.ac.nz

Supplementary information: http://biochem.otago.ac.nz/double.html

1 INTRODUCTION

Overlapping gene coding sequences (CDSs) occur where a nucleotide sequence codes for two different amino acid sequences in different read-frames. There are five possible read-frames for a 'secondary' open reading frame (ORF) relative to a 'primary' ORF that we label −3, −2, −1, +1 and +2 (Fig. 1). In terms of nucleotide mutation patterns, +1 and +2 are equivalent upon interchanging the primary and secondary ORFs.

Overlapping CDSs occur frequently in viruses (Normark *et al*., 1983), where they serve as a mechanism for fitting more genetic information into a small genome and for coregulating gene expression. In the hepatitis B virus (HBV), for example, ∼50% of the genome comprises overlapping CDSs (Mizokami *et al*., 1997). Overlapping CDSs also occur in prokaryotes (Rogozin *et al*., 2002; Fukuda *et al*., 2003) and, more rarely, in eukaryotes (Sharpless and DePinho, 1999; Poulin *et al*., 2003). Ribosomal frameshifting (Farabaugh, 1996) can also lead to partially overlapping CDSs in the +1 and +2 frames.

The majority of current gene-finding algorithms are optimized for finding non-overlapping protein-coding genes. Such methods are well-developed and generally make use of

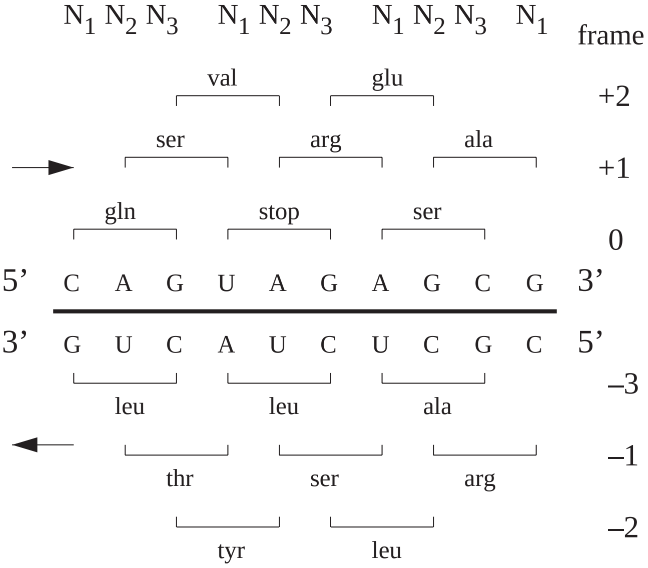
N1 N2 N3 N1 N2 N3 N1 N2 N3 N1
frame
val
glu
+2
ser
arg
ala
+1
gln
stop
ser
0
5' C A G U A G A G C G 3'
3' G U C A U C U C G C 5'
−3
leu
leu
ala
−1
thr
ser
arg
−2
tyr
leu

Fig. 1. Naming convention for the different possible read-frames of a secondary ORF relative to a frame = 0 primary ORF. In this paper, we use the terms 'primary' and 'secondary' to clarify which read-frame we are referring to; it does not imply that one ORF is more important than the other.

[*] To whom correspondence should be addressed.

combinations of the following signatures of protein-coding genes: (1) signal (e.g. splice sites, ORFs), (2) content (e.g. codon bias), (3) similarity to known sequences or conservation between species and (4) expression in cDNA/expressed sequence tag (EST) libraries (Stormo, 2000; Snyder and Gerstein, 2003).

Although standard gene-finding algorithms can also be used to find overlapping CDSs, there are a variety of problems which can lead to a decrease in sensitivity. Owing to the double-coding constraints, overlapping CDSs often display an atypical codon bias (Pavesi, 2000). Extending training set methods, such as hidden Markov models (HMMs), to overlapping CDSs is made difficult by several different frames (each requiring its own model) and limited training data. Similarity to known sequences or conservation between species may only point to the existence of one of an overlapping pair. Furthermore, overlapping genes on the same read-strand (e.g. at ribosomal frameshifting sites) may have the same promoter and mRNA, so that looking for promoters or expression may only identify one of the two genes.

Nonetheless overlapping CDSs have their own signatures resulting from the mutational constraints imposed by the requirement of simultaneously maintaining protein function in both genes. For example, previous studies have investigated (1) relative substitution rates in N_1, N_2 and N_3, i.e. the 1st, 2nd and 3rd nucleotide positions in each codon (Bilsel *et al*., 1990); (2) information theory indices (Pavesi *et al*., 1997); (3) low rate of synonymous mutations relative to other sites (Mizokami *et al*., 1997; Pavesi, 2000); and (4) codon usage (Pavesi *et al*., 1997; Pavesi, 2000). Perhaps the simplest method is to measure the mutation frequency in N_3 relative to N_1 and N_2. In single-coding sequences, many N_3 mutations give rise to synonymous amino acids and so N_3 is relatively unconstrained. In contrast, in double-coding sequences (except the -2 frame) N_3 is generally constrained since it corresponds to N_1 or N_2 in the alternative frame.

Given a known CDS in a sequence alignment and a potential overlapping CDS (i.e. a conserved or semi-conserved ORF), one may proceed to measure the conservation in N_1, N_2, N_3 within the overlap region and compare it with the background (i.e. single-coding) statistics measured in the non-overlap region of the known CDS. However, it is not clear how the results should be interpreted. For example: How significant are any observed deviations? What are the dependences on sequence divergence, composition and double-coding frame? What if the non-overlap region of the known CDS is too short to calculate background N_1, N_2, N_3 statistics?

In this paper, we investigate several simple statistics that may be used on pairwise alignments, namely the mutation rate in N_1, N_2, N_3 (N123) and the rate of synonymous and non-synonymous mutations (NsNn). We also develop a more sensitive and less frame-dependent, maximum-likelihood statistic (MLOGD) that makes use of a nucleotide substitution matrix, a codon usage table (CUT) and an amino acid substitution matrix. We develop a Monte Carlo sequence evolution algorithm that can produce simulated sequence alignments subject to either single-coding or double-coding constraints. Then we use the simulated sequences to characterize the distributions of the N123, NsNn and MLOGD statistics, as a function of sequence divergence, composition, length and frame, for both single-coding and double-coding sequences. By comparing observations for a real sequence alignment with such simulations, we may attempt to determine whether a potential overlapping CDS is real (i.e. functionally constrained) or not. We find that the new MLOGD statistic, $\Delta_{\mathrm{s,d}}$, is the most sensitive statistic for detecting overlapping CDSs.

2 METHODS

In Section 2.1, we describe the various statistics that we use to test for double-coding. In Section 2.2, we describe the Monte Carlo sequence evolution algorithm. In Section 2.3, we describe how we combine the two to classify a given sequence alignment as single-coding or double-coding. Further details are given in the Supplementary Material.

2.1 Double-coding test statistics

We investigate three test statistics designed to detect the mutation signature of overlapping CDSs in pairwise alignments of two sequences, S_1 and S_2.

2.1.1 N123: mutation rate in N_1, N_2, N_3 For the N123 method, we simply count the number of nucleotide differences between S_1 and S_2 in each of N_1, N_2 and N_3 (i.e. the 1st, 2nd and 3rd nucleotide positions in codons in the primary read-frame) and express each count as a fraction of the total number of N_1, N_2 and N_3 loci. We label these statistics f_{N1}, f_{N2} and f_{N3}, respectively.

2.1.2 NsNn: Synonymous and nonsynonymous mutation rates For the NsNn method, we step through aligned codon pairs in the primary read-frame of S_1 and S_2 and count which codon pairs are identical, non-identical but synonymous and non-synonymous. The numbers of synonymous and non-synonymous codon pairs are expressed as a fraction of the total number of codon pairs. We label these statistics f_{syn} and f_{non}, respectively.

2.1.3 MLOGD: maximum-likelihood method The MLOGD (Maximum -Likelihood Overlapping Gene Detector) method is an attempt to estimate the relative probabilities of S_1 mutating to S_2 under single-coding and double-coding models.

The evolution of a single-coding sequence is often modelled as a Markov process (Goldman and Yang, 1994). The probability of S_1 mutating to S_2 after time t may be expressed as

$$\log P(S_1 \to S_2; t) = \sum_{k=1}^{N_{\mathrm{codons}}} \log p_{C_1^k C_2^k}(t), \qquad (1)$$

where C_1^k, C_2^k are the k-th codons of S_1, S_2 and $[p_{ij}(t)] = \mathbf{P}(t) = \exp(\mathbf{Q}t)$, where $\mathbf{Q}$ is a 64×64 matrix of 'instantaneous' codon mutation probabilities.

In the case of overlapping CDSs, cross-talk between adjacent codons prevents the factorization-by-codon seen in (1). In theory, instead of a 64×64 codon matrix $\mathbf{Q}$, we must define a single $4^N \times 4^N$ matrix (where N is the sequence length in nucleotides) that describes the entire sequence at once. Clearly, for typical values of N, this is computationally impractical. Instead, we use the following simplified approach.

For each nucleotide pair N_1^k, N_2^k in S_1, S_2, we estimate the probability that N_1^k mutates to N_2^k for each of the single-coding and double-coding models. Specifically, we first define $b(N_1^k \to i; t, m)$, $i = \text{U, C, A, G}$, by

$$b(N_1^k \to i; t, \text{s}) = \mathbf{P}(N_1^k \to i; t) \times \mathbf{C}(X_2) \times \mathbf{A}(X_1 \to X_2) \qquad (2)$$

for the single-coding ($m =$'s') model, and

$$b(N_1^k \to i; t, \text{d}) = \mathbf{P}(N_1^k \to i; t) \times \mathbf{C}(X_2) \times \mathbf{A}(X_1 \to X_2) \times \mathbf{C}(X_2') \times \mathbf{A}(X_1' \to X_2') \qquad (3)$$

for the double-coding ($m =$ 'd') model. Here X_1 and X_2 are the original and final amino acids or codons in the primary read-frame and X_1' and X_2' are the original and final amino acids or codons in the secondary read-frame for the nucleotide mutation $N_1^k \to i$. Also $\mathbf{P}(t) = \exp(\mathbf{Q}t)$, and $\mathbf{Q}$, $\mathbf{C}$ and $\mathbf{A}$ are nucleotide, codon and amino acid substitution matrices, respectively (described in Section 2.2).

The probability that N_1^k mutates to N_2^k, after time t, is then given by

$$P(N_1^k \to N_2^k; t, m) = \frac{b(N_1^k \to N_2^k; t, m)}{\sum_{i=\text{U,C,A,G}} b(N_1^k \to i; t, m)}. \qquad (4)$$

As in (1), we sum over the sequence alignment as follows:

$$\log P(S_1 \to S_2; t, m) = \sum_{k=1}^{N_{\text{nucleotides}}} \log P(N_1^k \to N_2^k; t, m). \qquad (5)$$

We maximize (5) with respect to t for each of the single-coding and double-coding models, giving t^{s} and t^{d}, respectively. Then the log-likelihood ratio of the two models is

$$\Delta_{\text{s,d}} = \log P(S_1 \to S_2; t^{\text{d}}, \text{d}) - \log P(S_1 \to S_2; t^{\text{s}}, \text{s}). \qquad (6)$$

If $\Delta_{\text{s,d}}$ is positive, then the observed mutations between S_1 and S_2 are more consistent with double-coding. If $\Delta_{\text{s,d}}$ is negative, then the observed mutations are more consistent with single-coding.

The above methodology involves several approximations. For each component of the probability sum (5), we consider only a single nucleotide and the single codon in each of the primary and secondary read-frames containing that nucleotide; longer-range dependences are ignored. In addition, as far as codon and amino acid weightings are concerned, we consider only the start and end points of the unknown mutation pathway connecting an aligned codon pair in S_1 and S_2. It seems reasonable that these simplifications are justified provided S_1 and S_2 are not too divergent (so that mutation pathways are short and inter-codon cross-talk is low). Tests with simulated sequences (which are not subject to these simplifications) show that the model provides useful results over a wide range of circumstances (see Sections 3.1.1–3.1.4).

2.2 Monte Carlo sequence evolution model

The problems outlined above do not arise when generating simulated sequences. Random nucleotide mutations may be applied sequentially and the neighbourhood of any mutating nucleotide is known at the time of mutation, thus allowing appropriate amino acid and codon substitution weights to be applied.

Given a starting sequence, a frame, a mutation rate, a 4×4 nucleotide mutation matrix $\mathbf{Q}$, a 64-entry CUT $\mathbf{C}$ and a 20×20 amino acid substitution matrix $\mathbf{A}$, our Monte Carlo simulation applies nucleotide mutations one-by-one until the required total number of mutations have occurred. Nucleotide mutations are chosen randomly with probabilities determined by $\mathbf{Q}$ and are accepted or discarded with probabilities determined by $\mathbf{C}$ and $\mathbf{A}$ (for full details see Supplementary Material). Under the double-coding model, the codon and amino acid weights are applied in both read-frames. In order to match the simulations to real sequence data, the mutation rate λ is tied to the number of observed mutations rather than the number of accepted mutations (e.g. A → C → G counts as two accepted mutations but only one observed mutation).

By default, we use a $\kappa = 3$ Kimura (1980) nucleotide matrix, null CUT (i.e. equal codon frequencies) and the Henikoff and Henikoff (1992) BLOSUM40 amino acid substitution matrix (for details see Supplementary Material).

2.3 Classification

Having calculated the statistics $\Delta_{\text{s,d}}$, f_{N1}, f_{N2}, f_{N3}, f_{syn}, f_{non} for a particular pairwise sequence alignment $S_1 + S_2$, as described in Section 2.1, we then wish to classify it as either single-coding or double-coding according to which of the two models is most consistent with the observed statistics. For the MLOGD method, classification is straightforward: if $\Delta_{\text{s,d}} > 0$ we classify $S_1 + S_2$ as double-coding, while if $\Delta_{\text{s,d}} \leq 0$ we classify $S_1 + S_2$ as single-coding. For the N123 and NsNn methods, we use the simulations to find out what range of values we would expect to observe for the two models. We may then find the log-likelihood ratios

$$\Delta_{\text{s,d}}^{\text{N123}} = \log\left[\frac{P(f_{\text{N1}}, f_{\text{N2}}, f_{\text{N3}}|\text{d}, \lambda)}{P(f_{\text{N1}}, f_{\text{N2}}, f_{\text{N3}}|\text{s}, \lambda)}\right] \qquad (7)$$

and

$$\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}} = \log\left[\frac{P(f_{\mathrm{syn}}, f_{\mathrm{non}}|\mathrm{d}, \lambda)}{P(f_{\mathrm{syn}}, f_{\mathrm{non}}|\mathrm{s}, \lambda)}\right]. \tag{8}$$

We generate simulated sequences (Section 2.2), starting with the sequence S_1 and with the mutation rate fixed by the total number of point differences between S_1 and S_2. We generate 100 simulated sequences each for the single-coding and double-coding models. We then calculate f_{N1}, f_{N2}, f_{syn}, f_{non} for each of the 200 sequence pairs comprising S_1 and one of the simulated sequences, and calculate the means and SD, μ_x^m, σ_x^m, where m = 's' or 'd' is the model, and x is one of the statistics f_{N1}, f_{N2}, f_{syn}, f_{non}. Note that we omit f_{N3} since, having fixed $\lambda = f_{\mathrm{N1}} + f_{\mathrm{N2}} + f_{\mathrm{N3}}$, f_{N3} adds no new information.

Assuming normal distributions and assuming independence of f_{N1}, f_{N2} and of f_{syn}, f_{non}, it is straightforward to calculate $\Delta_{\mathrm{s,d}}^{\mathrm{N123}}$ and $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}}$ (see Supplementary Material). Since, in fact, these two assumptions are only approximations, $\Delta_{\mathrm{s,d}}^{\mathrm{N123}}$ and $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}}$ are not strictly speaking likelihood ratios, but may still be used as classifiers: the sequence pair, $S_1 + S_2$, is classified as single-coding if $\Delta_{\mathrm{s,d}}^{\mathrm{N123}} \leq 0$ or $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}} \leq 0$ and as double-coding if $\Delta_{\mathrm{s,d}}^{\mathrm{N123}} > 0$ or $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}} > 0$. The magnitudes of $\Delta_{\mathrm{s,d}}$, $\Delta_{\mathrm{s,d}}^{\mathrm{N123}}$ and $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}}$ also give a measure of the confidence of classifications (see Supplementary Material).

3 RESULTS

In Section 3.1, we use simulated sequences to characterize the statistics introduced in Section 2.1. We investigate their dependence on sequence divergence, double-coding frame (Section 3.1.1) and sequence length (Section 3.1.2). We also investigate how robust the statistics are with respect to choice of input nucleotide, codon and amino acid matrices (Section 3.1.3) and with respect to sequencing errors (Section 3.1.4). In Sections 3.2 and 3.3, we test the algorithms on real sequence data.

3.1 Tests on Monte Carlo simulations

3.1.1 Dependence on frame Simulated sequences (Section 2.2) were generated from initial random nucleotide sequences with a range of mutation rates $0 \leq \lambda \leq 0.5$, where λ is the number of observed mutations per nucleotide. Each simulated sequence S_2, together with the corresponding initial sequence S_1, makes an aligned sequence pair for which the statistics $\Delta_{\mathrm{s,d}}$, f_{N1}, f_{N2}, f_{syn}, f_{non} of Section 2.1 may be calculated.

Figure 2 shows the distribution of the MLOGD statistic, $\Delta_{\mathrm{s,d}}$, as a function of λ, for the different frames. MLOGD clearly separates the single-coding from the double-coding simulations, except at the smallest λ values (see Section 3.1.2), irrespective of frame. Figures 3 and 4 show the distribution of the N123 and NsNn statistics, f_{N1}, f_{N2}, f_{syn}, f_{non}, for the different frames. Here the distinction between single-coding and double-coding is often less clear, and much more frame-dependent.

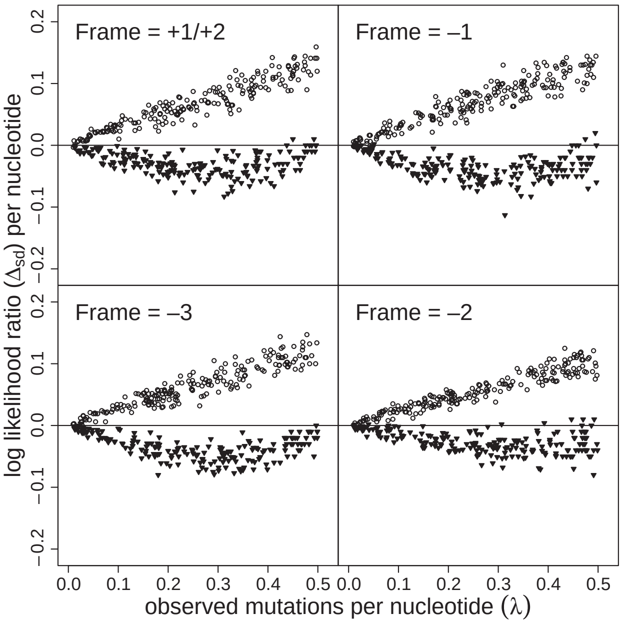

Fig. 2. Plots of the MLOGD statistic $\Delta_{\mathrm{s,d}}$ (per nucleotide) for simulated single-coding (triangles) and double-coding (circles) sequences of length 300 codons. Each symbol represents one sequence pair, $S_1 + S_2$, with the divergence between S_1 and S_2 measured on the x-axis. The four panels show the distribution of $\Delta_{\mathrm{s,d}}$ values for single-coding sequences and the five possible read-frames of an overlapping secondary ORF relative to a primary ORF. $\Delta_{\mathrm{s,d}} < 0 \Rightarrow$ single-coding while $\Delta_{\mathrm{s,d}} > 0 \Rightarrow$ double-coding. In contrast to Figures 3 and 4, where the single-coding f_{N1}, f_{N2}, f_{syn}, f_{non} values can be displayed in a separate panel, the single-coding $\Delta_{\mathrm{s,d}}$ values depend on which potential secondary read-frame the single-coding model is being tested against.

3.1.2 Dependence on sequence length It is apparent that the ability of any method to distinguish single-coding from double-coding regions will decrease for lower mutation rates λ and for shorter sequence lengths. Figures analogous to Figures 2–4 for sequence lengths 300, 50 and 20 codons are given in the Supplementary Material. Figure 5 shows the variation in the spread of $\Delta_{\mathrm{s,d}}$ values for sequence lengths 1000, 300, 50 and 20 codons for the +2 frame (e.g. the frame of the P, C and X genes relative to the S, P and P genes, respectively, in HBV and also the frame of −1 ribosomal frameshifts). The longest overlap in HBV is ~440 codons, though overlap regions are generally much shorter. Even for a sequence length of 20 codons, $\Delta_{\mathrm{s,d}}$ performs fairly well for $\lambda \geq 0.2$. In contrast f_{syn}, f_{non} and, to some extent, f_{N1}, f_{N2}, are much less useful for such short sequences (see below for details).

It is difficult to tell from such plots which statistics are most useful—especially for small λ, where identifying

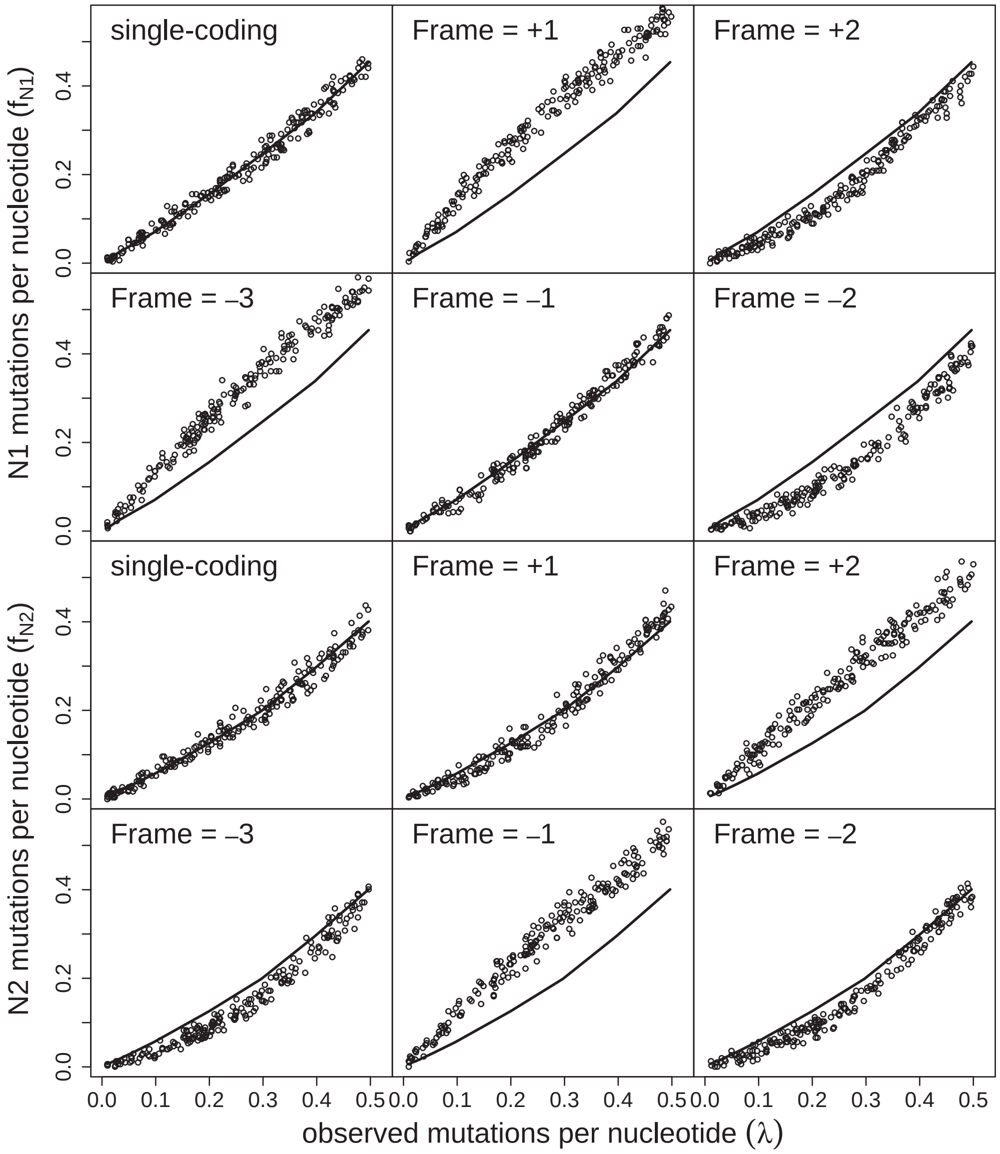

Fig. 3. Plots of the 1st and 2nd codon position mutation rates, f_{N1} and f_{N2}, as a function of λ for simulated sequences of length 300 codons. Each symbol represents one sequence pair, $S_1 + S_2$, with the divergence between S_1 and S_2 measured on the x-axis. The 12 panels show the f_{N1} and f_{N2} values for single-coding sequences and the five possible read-frames of an overlapping secondary ORF relative to a primary ORF. The line—marking the median values for single-coding sequences (first panel)—is included in each plot for reference. Note that the plots are highly frame-dependent and in some frames f_{N1} or f_{N2} taken alone is insufficient to distinguish double-coding from single-coding sequences.

double-coding regions is most challenging. To more precisely quantify the utility of the three methods of Section 2, we use the procedure described in Section 2.3 to classify each of 100 single-coding and 100 double-coding simulations (for a particular initial sequence, frame and mutation rate λ), calculating μ_x^m, σ_x^m, from the other 99 single-coding and 99 double-coding simulations. The number of single-coding simulations correctly classified as single-coding and the number of double-coding simulations correctly classified as double-coding give a measure of the power of each method.

These values are summarized in Figure 6 for the +2 frame (see Supplementary Material for other frames). For long sequences, MLOGD and N123 give good results. For short sequences, NsNn becomes unusable. The MLOGD statistic, $\Delta_{s,d}$, is consistently the most discriminating classifier, especially for small λ. For a sequence of 300 codons, MLOGD has a mean success rate (averaged over all five frames) of 97% for λ as low as 0.03 (cf. 74% for NsNn and 88% for N123). For a sequence of 20 codons, MLOGD has a mean success rate of 83% for $\lambda = 0.2$ (cf. 66% for NsNn and 77%

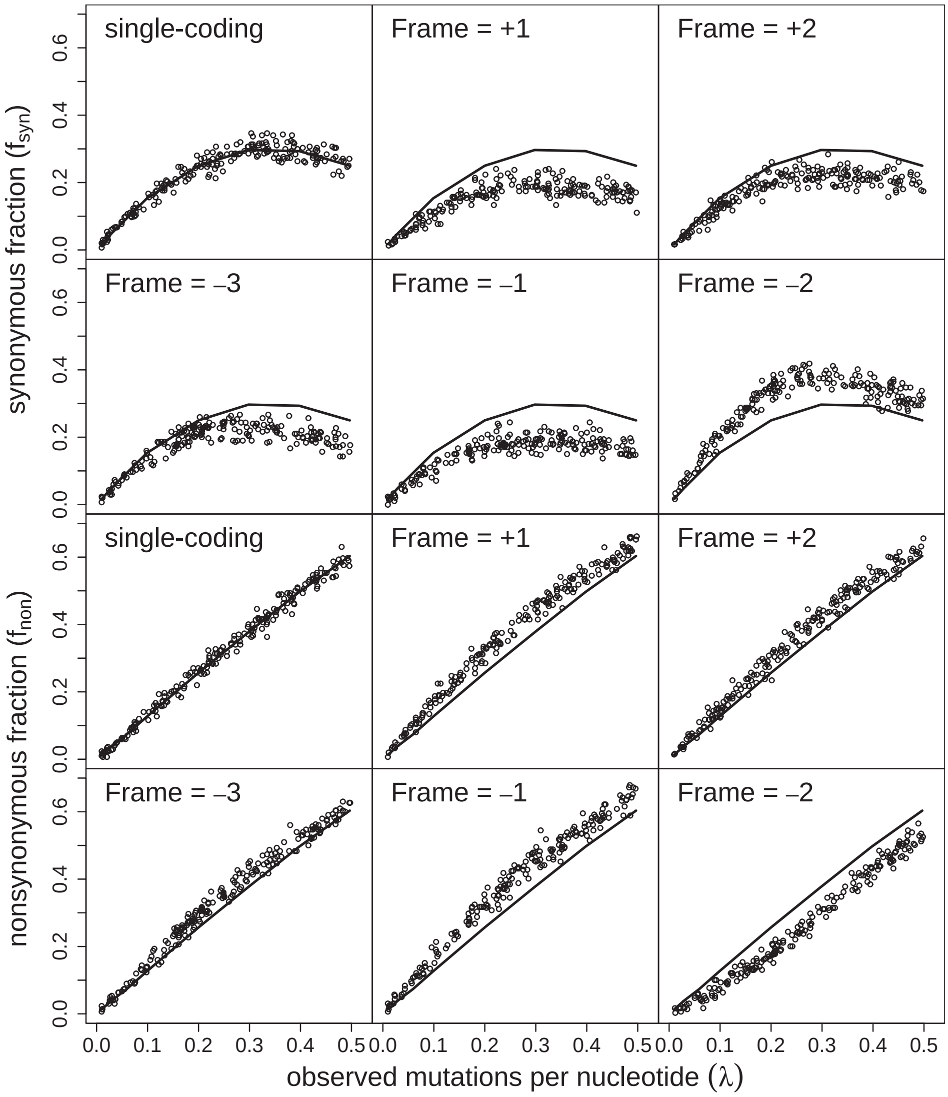

Fig. 4. Plots of the synonymous and non-synonymous mutation rates, f_{syn} and f_{non}, for simulated sequences of length 300 codons, double-coding in different frames (see caption to Fig. 3 for details). Note that in some frames it is often difficult to distinguish double-coding from single-coding sequences on the basis of f_{syn} and f_{non}. The -2 frame stands out as favouring synonymous mutations relative to non-synonymous mutations even more than single-coding regions. This is because in the -2 frame, N_3 in the primary CDS is opposite N_3 in the secondary CDS so this position is relatively unconstrained, whereas N_1 and N_2 are opposite N_2 and N_1, respectively, so these positions are highly constrained.

for N123). The NsNn results are little better than a random classifier (success rate 50%). These results are summarized in Figure 7.

3.1.3 Effect of model parameters We tested how robust the algorithms are with respect to the choice of input nucleotide, codon and amino acid matrices (Section 2.2). Simulated sequences were generated with (1) the default 50% GC-content nucleotide matrix replaced with a 70% GC matrix, (2) the default null CUT replaced with a human CUT or (3) the BLOSUM40 amino acid matrix replaced with the BLOSUM62 matrix. These sequences were classified as in Section 3.1.2, with the μ_x^m, σ_x^m model values derived from simulated sequences using the default matrices. Figures analogous to Figure 6 for (1), (2) and (3) are available in the Supplementary Material. MLOGD, NsNn and N123 are all robust with respect to reasonable changes in the model matrices.

3.1.4 Effect of sequencing errors We also tested how robust the algorithms are with respect to sequencing errors. Random nucleotide substitution errors were added to

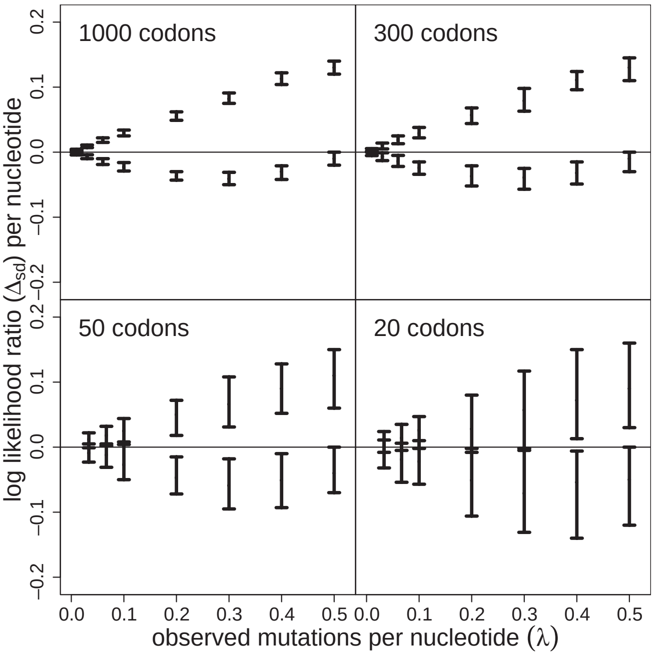

Fig. 5. Range of the MLOGD statistic $\Delta_{s,d}$ (per nucleotide) for simulated single-coding (lower bars) and double-coding in the +2 frame (upper bars) for sequences of various lengths. See caption to Figure 2 for details. The bars represent the central 68% of values (i.e. $\pm 1\sigma$ for normal distributions) based on 100 simulations at each of λ = 0.01, 0.03, 0.06, 0.1, 0.2, 0.3, 0.4 and 0.5. Even for overlap regions as short as 20 codons, MLOGD can often successfully distinguish double-coding from single-coding.

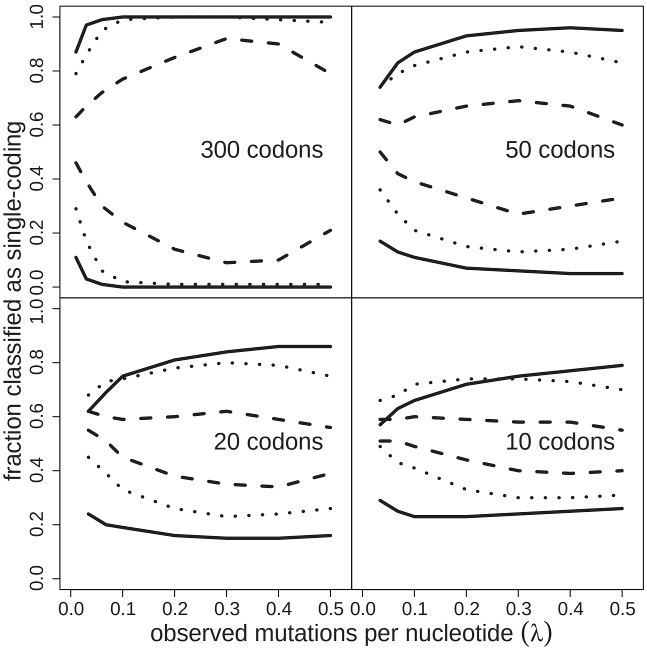

Fig. 6. Classification (Section 2.3) success as a function of mutation rate λ and sequence length, for the +2 frame. In each panel, the upper three lines show the fraction of single-coding simulations classified as single-coding (1 ≡ perfect classification) while the lower three lines show the fraction of double-coding simulations classified as single-coding (0 ≡ perfect classification). Overall classification success depends on the distance between the upper and lower lines. The classifications use the MLOGD (solid lines), NsNn (dashed lines) and N123 (dotted lines) methods (Section 2.1). Classification success is reduced for short sequences and low λ. MLOGD consistently gives the best classification. (Plotted points are averages for $\sim \frac{2000}{N_{\text{codons}}}$ random initial sequences. The rms error is of the order 0.01–0.03.)

simulated sequences at a rate of 0.3, 1 and 3% and the resulting sequences were classified as in Section 3.1.2, with the μ_x^m, σ_x^m model values derived from simulated sequences without errors. Figures analogous to Figure 6 for the different error rates are available in the Supplementary Material. MLOGD, NsNn and N123 are all similarly affected by sequencing errors. Averaging over all five frames, the decrease in classification success is about $0.3 \times \frac{\text{error rate}}{\text{mutation rate}}$ (e.g. for $\lambda = 0.1$ and a 1% sequencing error rate, the decrease in classification success is typically 3%). When the error rate is of the order of a third of the mutation rate, or greater, all methods are significantly affected.

3.2 Tests on hepatitis B (HBV) genome

We have tested the algorithms on known overlapping CDSs and non-coding overlapping ORFs in the HBV genome. HBV has a circular partially double-stranded DNA genome. The long strand comprises ~3215 nt and encodes four genes (P, C, S, X) read in the forward direction. The S gene is completely contained in the P gene and read in the +1 frame relative to P, while the C and X genes both overlap the ends of the P gene and are read in the +2 frame (Fig. 8).

We used 10 human strains (GenBank accession numbers NC_003977, X70185, D00329, X75665, AB074755, D50520, X02496, X75664, X75663 and AB056514), illustrating a range of diversity [genotypes C2, A, B, C, C1, C2, D, E, F and G; see Huy *et al*. (2004)], together with strains from woolly monkey (AF046996) and woodchuck (J02442). A phylogenetic tree is given in the Supplementary Material.

Multiple sequence alignments were made separately for each of the P, C, S, X ORFs using CLUSTALW (Higgins *et al*., 1994) on the translated sequences. In the interests of demonstrating a fully automated approach, no manual adjustments were made to the alignments. ORFs were detected in the reference sequence NC_003977 using the EMBOSS program getorf (Rice *et al*., 2000). For each potential overlap region in the reference sequence, MLOGD, NsNn and N123 statistics were calculated for each of the other 11 sequences paired with the reference sequence.

3.2.1 Hepatitis B—known overlaps The results for the overlap between the S and P genes are shown in Figure 9. This overlap region contains ~440 codons, providing a good signal for all three methods. The overlaps between C and

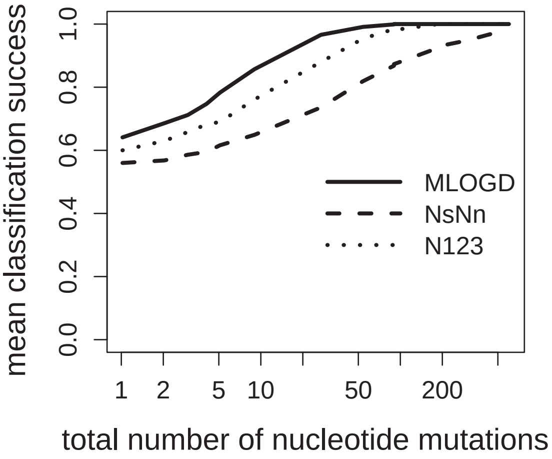

Fig. 7. Mean classification (Section 2.3) success for single-coding and double-coding sequences, as a function of the total number of point nucleotide differences between pairs of sequences (e.g. $\lambda = 0.2$, $N_{\text{codons}} = 20$ and $\lambda = 0.01$, $N_{\text{codons}} = 400$ are equivalent here). The results are averaged over all five frames.

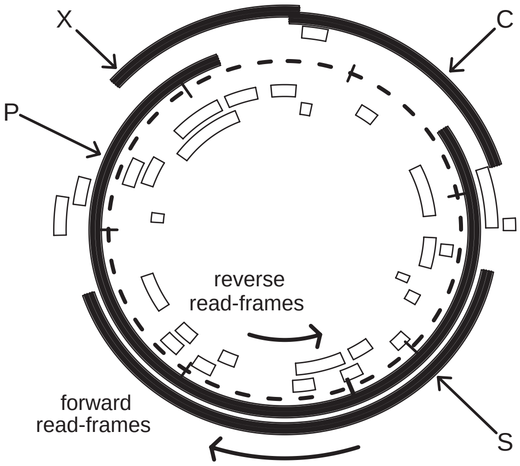

Fig. 8. Diagram of the circular HBV genome. P, C, S and X are known protein-coding genes. Open boxes represent all other ORFs in the reference sequence, NC_003977, that are 30 nt or longer. The radial tickmarks on the dashed centre-line are at 500-nt intervals.

P and between X and P are shorter (∼45 and ∼80 codons, respectively) and the signal is correspondingly reduced. All three methods give incorrect classifications for some sequence pairs, with the fraction of sequence pairs correctly classified being 86, 59 and 77% for MLOGD, NsNn and N123, respectively (see Supplementary Material). For MLOGD and N123, nearly all the incorrect classifications have low confidence (i.e. scores close to zero).

Note that the observed statistics are not infrequently outside the confidence limits predicted by the simulations. One reason for this is that the mutational patterns in the compact genomes of viruses are subject to many other constraints that can produce deviations from the models.

Tests with known overlapping CDSs in Lentiviruses, Luteoviruses and Poleroviruses produced similar results (Firth A.E. and Brown C.M., manuscript in preparation).

3.2.2 Hepatitis B—potential overlaps We also tested the algorithms on all other ORFs in NC_003977 that have at least 30-codon overlaps with one or more of the known P, C, S, X genes (Fig. 8). There are 43 such overlaps, involving 29 new ORFs. These ORFs are presumed to be non-coding, however short functional ORFs have been found in other viruses. Unlike the four known genes, many of these ORFs are in the reverse read-direction. Some are conserved in only a few strains and, although we skip stop ↔ non-stop codon transitions in the calculation of the MLOGD statistic, their presence in an alignment would commonly be taken as an evidence that an ORF is non-functional. Figures analogous to Figure 9 for all 43 overlaps, together with a summary table of statistics, are given in the Supplementary Material.

When scores are summed over all 11 sequence pairs, MLOGD, NsNn and N123 identify 79, 77 and 74%, respectively of the 43 overlaps as single-coding. One reason for the ∼21–26% failures may be due to the short length of some of the overlap regions—giving a reduced signal. Another important factor is that in viruses there are often several constraints, besides maintaining protein function, that contribute to the pattern of sequence conservation at any site. In particular, a large number of these ORFs overlap two known genes (e.g. S and P) thus giving an extra frame-dependent constraint on N_1, N_2, N_3, which essentially invalidates the N123 method. The MLOGD method is more robust with respect to this effect and in fact there are only a few (namely 4 out of 20) particular combinations of frames—easily determined from simulations (see Supplementary Material)—that are liable to give rise to a false double-coding signal in a tertiary ORF overlapping two coding sequences. In fact such a combination of frames explains seven of the nine false positives identified by the MLOGD method. These false positives are easily removed since all seven ORFs give a negative signal relative to the other of the two overlapping known genes. Taking such double-overlaps into account, MLOGD has a 93% success rate on the 29 (presumed) non-functional ORFs, many of which are very short.

3.3 Tests on bacterial genomes

The much larger genomes of bacteria are not subject to the same size constraints as virus genomes, and the vast majority of short ORFs are expected to be non-coding. In addition, real overlaps may often be the result of random events (e.g. stop codon mutation) and may not be subjected to strong functional constraints (Fukuda *et al*., 2003). Hence, testing the algorithms on bacterial genomes can provide a good estimate of the false positive rate.

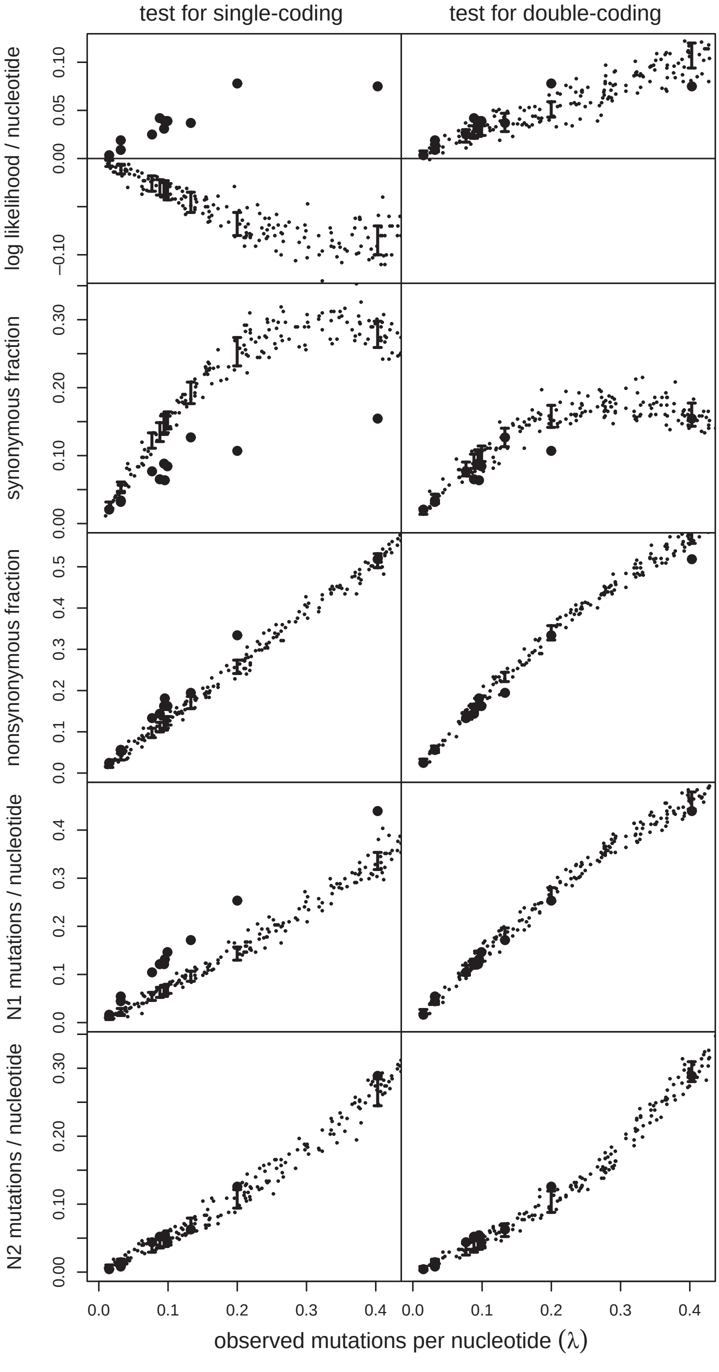

Fig. 9. The overlap between the S and P genes in HBV. Here P is taken as the primary ORF and S is in the +1 frame. The overlap region is ∼440 codons long, providing a good signal in all plots. The observed statistics for HBV pairwise alignments are represented by solid circles while the dots correspond to simulations—single-coding on the left and double-coding in the +1 frame on the right. Error bars are each based on 100 simulations with the same mutation rate λ as the corresponding sequence pair, and enclose the central 68% of values. MLOGD, NsNn and N123 all clearly identify this region as double-coding.

We used a table of 3682 pairwise symmetrical best hits between annotated protein-coding genes in *Escherichia coli* (K12; NC_000913.2) and *Salmonella typhimurium* (LT2; NC_003197.1), obtained from NCBI, as the set of primary ORFs. The set of secondary ORFs was taken to be all ORFs detected in *E.coli*. MLOGD scores were calculated (using the alignment between the primary ORFs in *E.coli* and *S.typhimurium* and an *E.coli* CUT) for all overlaps with length

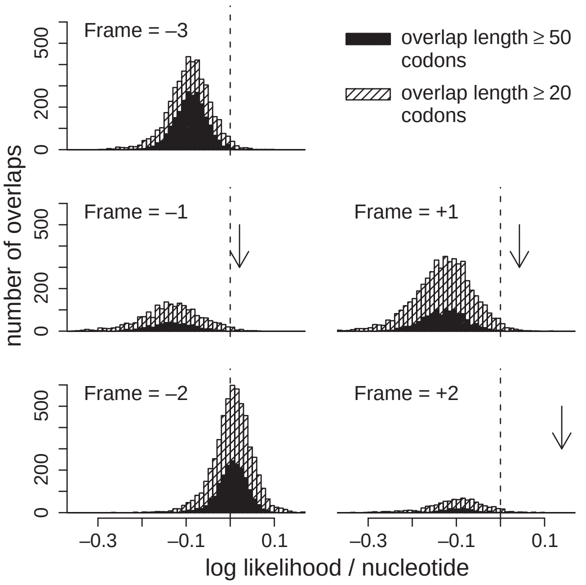

Fig. 10. Histograms of *E.coli–S.typhimurium* MLOGD scores (per nucleotide) for overlapping ORFs in *E.coli*. Scores less than zero (to the left of the dashed vertical line) imply single-coding, and are expected for the vast majority of overlapping ORFs. Arrows indicate the positions of the three annotated overlaps that are conserved in *S.typhimurium*. Except for the −2 frame, the false positive rate is very low.

at least 20 codons and sequence identity within the overlap of at least 70%. A total of 18 081 overlaps were analysed. Only 19 of these involved two annotated genes in *E.coli*, and only three of these pairs were conserved as an overlapping pair in *S.typhimurium*.

Figure 10 shows the distribution of MLOGD scores for the different frames. Except for the −2 frame, the number of MLOGD scores greater than zero is very low: 2 and 0.5% for overlap lengths ≥20 and 50 codons, respectively. Some of these overlaps may be functional—e.g. the three conserved annotated overlaps marked, plus possibly others which, even if not conserved in *S.typhimurium*, may still have been subject to double-coding constraints over part of the evolutionary divergence between the two species.

The false positive rate in the −2 frame is unacceptably high if a threshold score of zero is used. In this frame, N_3 in the primary ORF opposes N_3 in the secondary ORF, leaving N_3 relatively unconstrained. Thus evolution mimics evolution in single-coding sequences. Hence, a higher threshold needs to be used—e.g. using the variation in calculated MLOGD scores, as a function of overlap length, as a guide (see also Supplementary Material). Note also that overlaps in the −2 frame are expected to be relatively rare (Rogozin *et al*., 2002).

4 DISCUSSION

We have presented a new algorithm (MLOGD) for predicting whether or not a particular region in a pairwise sequence alignment is likely to be double-coding. We have also used simulations to formalize and put probabilities on two simple, but rather ill-defined, previously used test statistics (N123 and NsNn). All three methods have been fully characterized using simulated data, and tested on viral and bacterial genomes.

We find that no method gives perfect predictions all the time, but MLOGD consistently performs better than N123, while NsNn performs rather poorly. MLOGD is particularly useful when there are extra constraints on sequence evolution (e.g. known double-coding with a potential tertiary ORF, or known single-coding with two alternative potential secondary ORFs). MLOGD also has the advantage that the distribution of the $\Delta_{\mathrm{s,d}}$ statistic is relatively frame-independent and easy to interpret (namely, $\Delta_{\mathrm{s,d}} \leq 0 \Rightarrow$ single-coding and $\Delta_{\mathrm{s,d}} > 0 \Rightarrow$ double-coding). In contrast, the N123 and NsNn statistics require time-consuming simulations to interpret their values. Simulations suggest that MLOGD can identify double-coding regions at ~90% confidence for pairwise sequence alignments that differ at just ~20 nt sites (Fig. 7), though results with complex real sequence data are less good (possibly due to suboptimal alignments, Section 3.2).

Although MLOGD produced the best results, we suggest using all three methods and comparing results. By doing this, one can efficiently identify candidate overlapping CDSs for further investigation. In addition, comparison with the model simulations can provide valuable insights about the mutational constraints operating within a given sequence. For an input sequence alignment, our software package automatically produces plots similar to Figure 9 and a table of sequence statistics, allowing a quick visual inspection of double-coding constraints in any input primary ORF.

For many pairs of overlapping genes, one of the two is fairly unconstrained, e.g. in viruses many code for structural proteins whose primary sequence may not be highly conserved. In evolutionary terms, such a situation is desirable, since too many constraints can be a severe limitation on evolution (Rogozin *et al*., 2002). Our software also produces nucleotide-by-nucleotide plots for MLOGD. This allows one to quickly identify particularly conserved regions in an overlap. Such analyses can be useful for selecting targets for vaccines or antiviral drugs since, if functional constraints are important in both genes of an overlapping pair, then mutations will be more restricted.

Although the statistics $\Delta_{\mathrm{s,d}}$, $\Delta_{\mathrm{s,d}}^{\mathrm{N123}}$ and $\Delta_{\mathrm{s,d}}^{\mathrm{NsNn}}$ (Section 2.3) may be summed along the branches of a phylogenetic tree—for increased sensitivity—by calculating, for example, maximum parsimony sequences at intermediate nodes, we have preferred to display statistics for each pairwise alignment. When only a small numbers of species are available, it is informative to view each comparison individually. Even when large alignments are available, certain overlapping CDSs may exist in only some lineages, so again it is important to view the pairwise, rather than summed, statistics.

The algorithms should prove useful for detecting novel genes in viruses—where overlapping gene structures are common and often conserved over large evolutionary distances. They are expected to be particularly useful for analysing short ORFs and ribosomal frameshift sites. In addition, many overlapping CDSs have been annotated in yeast and prokaryotic genomes, and these algorithms could be used to test their functionality.

ACKNOWLEDGEMENTS

Thanks to Peter Stockwell, Amonida Zadissa and Rob Pollock for many interesting discussions. A.E.F. gratefully acknowledges funding from the Foundation for Research, Science and Technology, grant number UOOX0304.

REFERENCES

Bilsel,P.A., Rowe,J.E., Fitch,W.M. and Nichol,S.T. (1990) Phosphoprotein and nucleocapsid protein evolution of vesicular stomatitis virus New Jersey. *J.Virol.*, **64**, 2498–2504.

Farabaugh,P.J. (1996) Programmed translational frameshifting. *Annu. Rev. Genet.*, **30**, 507–528.

Fukuda,Y., Nakayama,Y. and Tomita,M. (2003) On dynamics of overlapping genes in bacterial genomes. *Gene*, **323**, 181–187.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

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

Huy,T.T., Ushijima,H., Quang,V.X., Win,K.M., Luengrojanakul,P., Kikuchi,K., Sata,T. and Abe,K. (2004) Genotype C of hepatitis B virus can be classified into at least two subgroups. *J. Gen. Virol.*, **85**, 283–292.

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

Mizokami,M., Orito,E., Ohba,K., Ikeo,K., Lau,J.Y. and Gojobori,T. (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.*, **44**, S83–S90.

Normark,S., Bergstrom,S., Edlund,T., Grundstrom,T., Jaurin,B., Lindberg,F.P. and Olsson,O. (1983) Overlapping genes. *Ann. Rev. Genet.*, **17**, 499–525.

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

Pavesi,A., Iaco,B., Granero,M.I. and Porati,A. (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J. Mol. Evol.*, **44**, 625–631.

Poulin,F., Brueschke,A. and Sonenberg,N. (2003) Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *J. Biol. Chem.*, **278**, 52290–52297.

Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Rogozin,I.B., Spiridonov,A.N., Sorokin,A.V., Wolf,Y.I., Jordan,I.K., Tatusov,R.L. and Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, **18**, 228–232.

Sharpless,N.E. and DePinho,R.A. (1999) The INK4A/ARF locus and its two gene products. *Curr. Opin. Genet. Dev.*, **9**, 22–30.

Snyder,M. and Gerstein,M. (2003) Defining genes in the genomics era. *Science*, **300**, 258–260.

Stormo,G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.*, **10**, 394–397.