Quantitative Sequence and Open Reading Frame Analysis based on Codon Bias

Susan Rainey
Department of Haematology, Queen's University Belfast
Belfast BT9 7AB United Kingdom

and

Joe Repka
Department of Mathematics, University of Toronto
Toronto, Ontario M5S 3G3 Canada

ABSTRACT

The frequencies with which the sixty-four codons occur in human coding DNA are known. If we assume that the codons occur randomly, subject only to these probabilities, then it is possible to predict trinucleotide frequencies in each of the five other reading frames. A model is developed for evaluating the extent to which a given sequence has trinucleotide frequencies compatible with coding DNA. This model is tested using known samples of coding DNA taken at random from GenBank, and good agreement is found. Practical and theoretical applications are discussed, including determination of coding open reading frames, evaluation of sequence data for frameshift mutations and examination of hypothetical genes.

Keywords: codon bias, theoretical model, open reading frame, ORF, frameshift mutations, hypothetical genes

INTRODUCTION

It has been recognized for many years that one of the distinctive features of coding DNA is the phenomenon of codon bias. Since there are $4^3 = 64$ nucleotide triplets to specify only twenty different amino acids, the genetic code is degenerate. Interestingly, the synonymous codons that specify a given amino acid do not appear equally often. Across species, the codon bias varies greatly [1]. Codon usage tables have been developed for many species, including humans, by examining the protein coding genes available in GenBank [2].

The basis for the evolution and persistence, as well as the possible roles of codon bias have been studied, but are not completely understood. One method to determine codon frequencies is simply to count the number of times each of the 64 codons appears in a given piece of in frame coding DNA, and compare these values on a codon by codon basis with the values expected from the codon frequency tables. However, this works well only for long pieces of DNA. Shorter pieces, for example the size of a typical exon (say, 100 codons), do not contain enough codons to provide an accurate comparison with each of the 64 expected codon frequencies. More refined methods of examining codon bias began in 1982, when Staden and McLachlan first examined codon bias in longer strings of DNA by measuring the strength of codon preferences in successive sequence "windows". A summary of this and related work on the use of codon bias to identify coding DNA appears in Fickett [3], and it will be compared in the discussion section with the work presented here.

In this paper, the idea of codon bias is further developed and expanded to analyze DNA sequences in ways that provide new and different information. There are six frames in which DNA can be read. We use the knowledge of codon frequencies in the correct reading frame in coding DNA to predict trinucleotide frequencies in each of the other five (non-coding) frames. We develop a mathematical model which predicts the probability with which a hypothetical piece of coding DNA can be expected to occur in its coding frame and in each of the other five frames. This model is tested against sequences obtained randomly from GenBank, and the observed values conform remarkably well to those predicted by the model. The simultaneous pattern of trinucleotide usage in the six possible reading frames of a sequence produces a "signature" characteristic of coding DNA in a particular species. Comparisons of data from a given DNA sequence with the expected signature provide more information about its codon bias profile than previous methods. An earlier version of this work appeared in [3]. A number of practical and theoretical applications of this model are presented and discussed here.

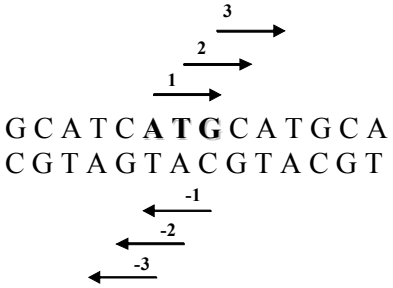
METHODS AND RESULTS

1. Codon bias and reading frames used in the model

For a given piece of DNA there are six possible reading frames, i.e., six ways in which the DNA can be grouped into trinucleotides (the true coding frame, if any, and five other non-coding frames). For the purposes of this study, when a sequence is known to be coding in a particular frame, the true coding frame will be referred to as reading frame 1, and the other five, non-coding frames as reading frames 2, 3, -1, -2, and -3, according to the scheme described in Figure 1(a). In this model, each of these six reading frames is analyzed for a given piece of DNA. When a sequence of DNA is presented, for example in a database like GenBank, it typically includes a number of bases upstream of any coding regions it may contain. The reading frame of a coding region is determined by locating the start of the coding region (for example, the ATG of the first exon) relative to the beginning of the actual sequence provided. It will depend on the number of upstream bases that have been included. For the purposes of notation in this study, this "nominal frame" of the given sequence will be called either frame I, II, III, -I, -II, or -III, according to the scheme described in Figure 1 (b).

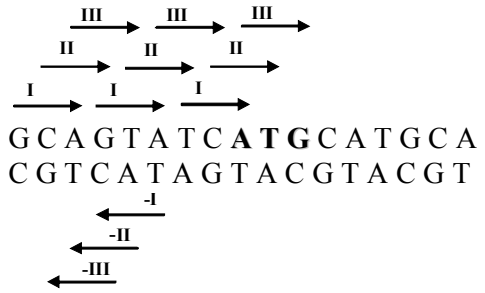
Figure 1 Frame conventions

(a)



For a given piece of DNA in any nominal frame, there are six possible reading frames, i.e., six ways in which the DNA can be grouped into trinucleotides. The convention used here is that reading frame 1 is the true coding frame, and reading frames 2, 3, -1, -2, and -3 represent the other five non-coding frames. A sample sequence is shown, with arrows and numbers indicating the position and direction of each reading frame. The trinucleotides in the true reading frame should thus follow the expected frequencies for coding DNA. Top strand = forward strand, bottom strand = reverse complement.

(b)



Convention used to describe the six nominal frames in DNA. The nominal frame is determined by locating the start of the coding region (here, ATG) relative to the beginning of the actual sequence provided. A sample sequence is shown, with arrows and numbers indicating the position and direction of each nominal frame. The example shown would be coding in nominal frame III, since the ATG begins at position III. Top strand = forward strand, bottom strand = reverse complement.

By examining sequences from GenBank, Nakamura *et al*. have produced codon frequency tables (available at http://www.kazusa.or.jp/codon/) for various species including humans. These frequencies refer to occurrences in reading frame 1. It is interesting to consider what the trinucleotide frequencies would be if such a sequence were read in one of the other five non-coding reading frames. Under the assumption that codons occur in random order, subject only to these frequencies, the frequencies of trinucleotides in the other five non-coding reading frames can be calculated as follows. For instance, for ACG to occur in frame 2, it must be constructed out of parts of two different frame 1 codons (*AC + G**). The four frame 1 codons of the form *AC have probabilities as follows: AAC = 0.0198, CAC = 0.0149, GAC = 0.0261, and TAC =0.0158. So the total probability that a frame 1 codon will end in AC is the sum of these probabilities, which is 0.0766. Similarly, the probability that a frame 1 codon will begin with G is the sum of the probabilities of the sixteen codons that begin with G, which is 0.3156. Thus the probability that ACG will occur in frame 2 is 0.0766 x 0.3156 = 0.0242. The other probabilities for trinucleotide frequencies in both frames 2 and 3 can be calculated analogously. For frames -1, -2, and -3, read on the reverse complement strand, the trinucleotide frequencies are calculated as follows. The probability that ACG will occur in frame -1 is equal to the probability that its reverse complement, CGT, will occur in frame 1. The probability that ACG will occur in frame -2 is equal to the probability that its reverse complement will occur in frame 3, and the probability that ACG will occur in frame -3 is equal to the probability that its reverse complement will occur in frame 2. The frequencies for all other trinucleotides can be found in this way (see Table 2).

2. Mathematical Model

We would like to determine, for a given sequence of length n codons, the probability that this sequence would arise if sequences of length n were generated at random subject to the known codon frequencies. Using the trinucleotide frequencies in all six reading frames, it is possible to predict not only the probability that the given sequence would arise at random in its coding frame, but also the probabilities that this same sequence would arise by chance in the other five non-coding frames. Let $p_i^{(k)}$ represent the probability of occurrence of the ith trinucleotide (out of the 64 possible codons) in frame k (i.e., one of the six possible reading frames) as calculated above. Consider a particular known segment of coding DNA made up of a string of n codons. Let n_i represent the number of times that the ith codon appears in the specified string in frame 1. Thus we can answer the question posed at the beginning of this section: the probability that the specified string would have occurred at random in frame k is the product of the probabilities of occurrence of each codon in the string, that is,

$$\prod_{i=1}^{64} (p_i^{(k)})^{n_i} . \qquad (1)$$

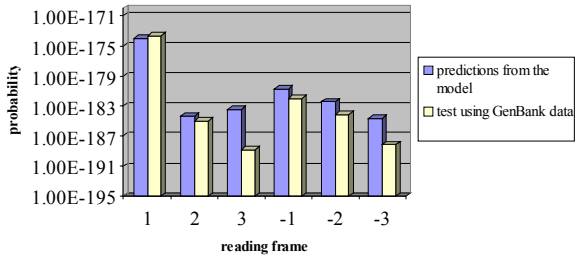
The second question we consider is the analogous problem, but this time for a typical but unspecified segment of coding DNA. Consider a segment of DNA coding in nominal frame I, made up of n codons. If the string of codons is long, (i.e., if n is large) it is reasonable to assume that the string will contain codons in the proportions given by the codon frequency database. Thus, the expected number of copies of the ith codon is $np_i^{(1)}$, where $p_i^{(k)}$ is defined as above. The probability that the unspecified string would occur at random in frame k (i.e., as a random string with the trinucleotide frequencies expected for frame k) is

$$\prod_{i-1}^{64} (p_i^{(k)})^{np_i^{(1)}} = \left(\prod_{i=1}^{64} (p_i^{(k)})^{p_i^{(1)}} \right)^n . \qquad (2)$$

Equation (2) was evaluated for each k, for a random string of n=100 codons (100 codons was chosen because it represents a typical exon length in human genes). The results are presented as the shaded bars in Figure 2. In the development of equation (2), the unspecified string of DNA was known to be coding in nominal frame I. It will also be useful to consider strings when it is not known in which nominal frame they are coding. If the calculation is performed for a string which in fact is coding in nominal frame III, then the probabilities obtained would be the same as those in Figure 2, but permuted so that the largest one was in the third position. From this we would infer that the true reading frame (reading frame 1, according to our convention) was nominal frame III. In this way, the probabilities calculated for sequences coding in the six nominal frames would produce six different graphs. The results show that a given string will have a higher probability of arising by chance in its correct reading frame, as opposed to the other five frames. Moreover, the probabilities that a given string will occur by chance in each

of the six frames provide a kind of numerical and graphical signature which indicates whether a particular string is likely to be coding DNA, and if so, in which frame.

Figure 2 Predictions and test of the mathematical model



The graph of predictions from the model (shaded bars) represents the expected probabilities in each of the six reading frames, of occurrence of a random string of DNA of 100 trinucleotides, conforming to the expected codon bias for frame 1. Note that the scale is logarithmic, indicating that the probability of the sequence occurring in frame 1 is approximately 10^7 times as large as the probability of this sequence occurring in the next most likely frame, which is frame -1.

The lighter bars represent the results based on seven human sequences, each coding in nominal frame I, which were randomly chosen from GenBank and broken down into 100 codon strings. Each string was evaluated to assess the probability that it would appear at random in each of the six reading frames. Of the total of fifty-four 100 codon strings tested, all clearly showed the highest probability of occurring in reading frame 1. Also, the pattern of probabilities observed across the six frames for each sequence clearly agreed with the results predicted by the model for a sequence coding in nominal frame I. The graph shows the geometric average, in each frame, of the probabilities of the fifty-four 100 codon strings tested. The accession numbers of the test sequences used were: L20046,M72393, L40157, D50063, M81695, L39068, and D00022.

3. Testing the model

The model was tested using a sample of human coding sequences randomly obtained from GenBank as follows. Using a random number generator to produce accession number prefixes, we searched for entries matching a particular prefix, which were also homo sapiens, complete coding sequences (cds), nuclear DNA (not mitochondrial), cDNA or RNA (no introns), and with an initiator ATG as the first codon (to clearly establish the reading frame). For any entry that fit these criteria, we broke the cds into consecutive strings of length 50 and 100 codons in nominal frame I, and calculated the probability that each would appear at random in each of the six frames, using equation (1)

$$\prod_{i=1}^{64} (p_i^{(k)})^{n_i}, \qquad (1)$$

where n_i is the number of times that the ith codon appears in the string (determined by counting the number of each of the 64 codons in the string). We did this until we had tested over one hundred 50-codon strings and over fifty 100-codon strings. For the one hundred and twelve 50-codon strings tested, all but seven had the highest probability of occurring in reading frame 1, and every one of the fifty-four 100-codon strings tested had highest probability of occurring in reading frame 1. Figure 2 shows the geometric average over the fifty-four 100-codon strings tested of the probabilities for each reading frame, compared with the results predicted by the model. Even with this relatively small sample, the agreement is apparent, confirming that the basic assumptions of our approach are not unreasonable.

4. Applying the Model to Sequence Analysis

4.1 Signatures for coding DNA in each frame This model can now be used to analyze any given sequence, to assess the probability that it is coding, and in which frame. The given sequence is read in each choice of nominal frame, and the probability of the resulting string occurring in each of the six reading frames is evaluated and compared to the probabilities predicted by the model for a sequence coding in that nominal frame. If this calculated signature shows similarity with the signature predicted by the model, then the sequence is likely to be coding in the chosen nominal frame. If the calculated signature does not show similarity with any of the signatures predicted by the model, then the sequence is likely not coding, as it does not conform to the codon bias expectations.

4.2 Weights of the signatures from each frame For a given string of DNA, the calculated signature is not likely to match exactly with any one of the six signatures predicted by the model. It would thus be useful to determine quantitatively the level of similarity of the calculated signature to each of the six predicted signatures, corresponding to the six possible nominal frames. To do this, we will write the calculated signature as a weighted sum of the predicted signatures.

For notational convenience, in this section only, we will label reading frames -1, -2, -3 as 4, 5, and 6. We will also label the nominal frames as 1, 2, 3, 4, 5, 6. We begin by expressing the calculated signature as a vector v, whose entries are the logarithms of the probabilities of occurrence of the string in the six reading frames. If p_i is the probability that the given string occurs in reading frame i, then

$$v = \begin{pmatrix} \log p_1 \\ \log p_2 \\ \log p_3 \\ \log p_4 \\ \log p_5 \\ \log p_6 \end{pmatrix}.$$

Now consider the six predicted signatures for a string of length n, the length of the given string. Let p_{ij} be the predicted probability that a typical coding sequence of length n in nominal frame j will occur in reading frame i. Each of the predicted signatures can be expressed as a vector v_j (j=1,2,...6), whose entries are the logarithms of the probabilities p_{ij} . The calculated signature vector v may now be expressed as a linear combination (weighted sum) of the predicted signature vectors: $v = c_1 v_1 + c_2 v_2 + c_3 v_3 + c_4 v_4 + c_5 v_5 + c_6 v_6$, where the numerical coefficients c_j can be interpreted as the weights contributed to the calculated signature by the predicted signatures. This amounts to the matrix equation v = A c, where $v = [\log p_i]$ is the vector containing the logarithms of the calculated probabilities for a given string, $A = [\log p_{ij}]$ is the matrix containing the logarithms of the predicted probabilities from the model, and $c = [c_j]$ is the vector containing the weights contributed by the predicted signatures to the observed signature. It can easily be established that the matrix A is invertible; denote the inverse matrix by A^{-1} . Thus, the matrix equation v = A c can be written as $A^{-1} v = A^{-1} A c$, so $c = A^{-1} v$. The values for c_j (j=1,2,...6) can easily be calculated and graphed, and the resulting plots are referred to as vector

coefficient graphs (examples appear in Figures 4). These vector coefficient graphs make it possible to visualize quickly the degree of similarity between the calculated signature and any of the predicted signatures. They also allow a rapid visualization of interesting patterns in the data.

4.3 Matching Scores The weights calculated in the previous section can now be used to determine if there is a good match between the calculated signature and any of the predicted signatures. Note that if one of the c_js equals 1 and the others are all 0, then the probabilities exactly match those predicted by the model for a sequence in frame j; this is of course unlikely. If one of the c_js is considerably larger than the others, it suggests that the sequence is close to what would be expected for a coding sequence in the corresponding frame. If none stands out, then the sequence is probably not part of a coding sequence.

We can determine a "matching score", by expressing the largest coefficient ($c_{j(L)}$ = largest c_j) as a proportion of the total length of vector c. The matching score is

$$s = \frac{c_{j_{(L)}}}{\sqrt{c_1^2 + c_2^2 + c_3^2 + c_4^2 + c_5^2 + c_6^2}}.$$

When the string under examination is expected to be coding in a particular frame, it is important to consider how well its vector coefficient graph fits the prediction from the model for that expected frame. If c_e is the coefficient corresponding to the expected frame, the matching score expresses c_e (instead of $c_{j(L)}$) as a proportion of the total length of vector c. This score indicates how well the signature for the particular string (e.g., an ORF) matches the signature preedicted for a string in its expected frame. The matching score for a specified expected frame e is

$$s = \frac{c_e}{\sqrt{c_1^2 + c_2^2 + c_3^2 + c_4^2 + c_5^2 + c_6^2}}.$$

Determining how well a given string fits the predictions of the model can be decided based on the value of the matching score. A perfect fit corresponds to a matching score of 1. For our purposes, we will consider a matching score $s < 0.50$ to be a negative fit, $0.50 \le s < 0.75$ to be inconclusive, and $0.75 \le s \le 1.00$ to be a positive fit. Note that it is possible to have negative scores.

5. Applications

The model presented here can now be further developed for use in practical and theoretical sequence analysis. In this section, three examples of sequence analyses that can be performed with the model are presented, and further potential applications being investigated are outlined in the discussion.

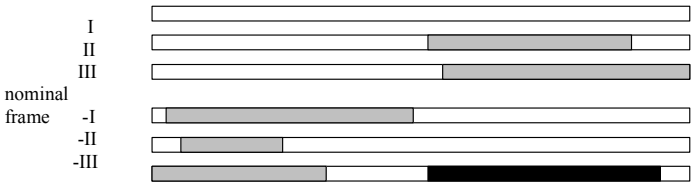
5.1 Evaluating open reading frames Any section of cDNA or mRNA can be searched to locate all possible open reading frames (ORFs), that is, regions displaying in frame start and stop codons. The ORF finder is a useful tool which is available through the National Center for Biotechnology Information (NCBI). It will evaluate a given sequence to determine ORFs in all possible frames. A typical search identifies many ORFs for a given sequence; however, usually only one is the "true" ORF, which is actually translated into a protein. Without experimental data, it can often be difficult to identify the true ORF.

The mathematical model presented here can be used to identify ORFs that seem, on the basis of codon bias, more likely candidates to be actual coding DNA. Each ORF is evaluated to determine the extent to which it does (or does not) conform to the predictions of the model, i.e., the extent to which it does (or does not) conform to expected patterns of codon bias. For example, suppose an ORF has been found in nominal frame I. If it codes for a protein, then the probabilities for each of the six reading frames should correspond to the signature expected for a piece of coding DNA in nominal frame I, as given in the model prediction. If these six probabilities do not correspond to the expected signature, then this ORF is probably spurious.

This is demonstrated with an example of a cDNA sequence obtained from GenBank, and tested with the NCBI ORF Finder and our model. The example is given in Figure 3, using the sequence under GenBank accession number BG056643. This sequence was derived from an EST and codes for at least part of an unknown protein. When this sequence is analyzed using the ORF Finder, seven different ORFs are found. It is impossible to predict from only the size and positions of these ORFs which one is actually a region that is translated into a protein. When each of these ORFs is tested using the model, only the ORF in nominal frame -III from bp 332 - 589 fits the expected signature for coding DNA. The results for the three largest ORFs are shown in Figure 3 (b) and (c). The vector coefficient graphs show that the calculated signatures for ORF 24 - 383 and ORF 22 - 378 do not match the predictions for any for the six nominal frames. Thus, these ORFs are likely not coding DNA segments. However, the graph of the calculated signature for ORF 332 to 589 (nominal frame -III) matches the predicted signature for nominal frame -III quite well. Thus, this ORF, shown shaded in Figure 3 (a), is predicted to be a region translated into a protein. More recent information added to GenBank has indicated that this EST is predicted to be similar to the insulin-like growth factor binding protein 2 precursor (GenBank accession number XM_002636) in this ATG-stop region and further upstream in nominal frame -III. Comparison of these gene sequences shows that this would mean that this EST is indeed coding in nominal frame -III, as predicted by the model.

Figure 3 Using the model to test ORFs

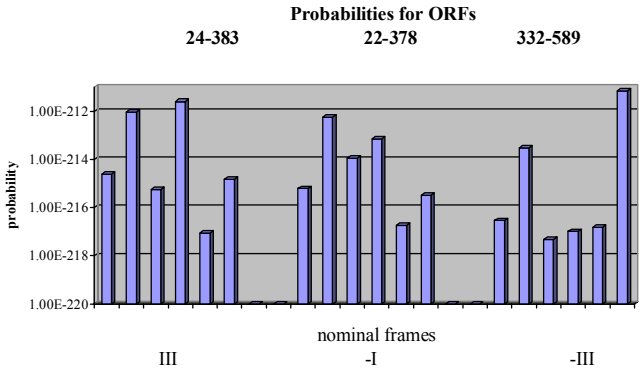
(a)



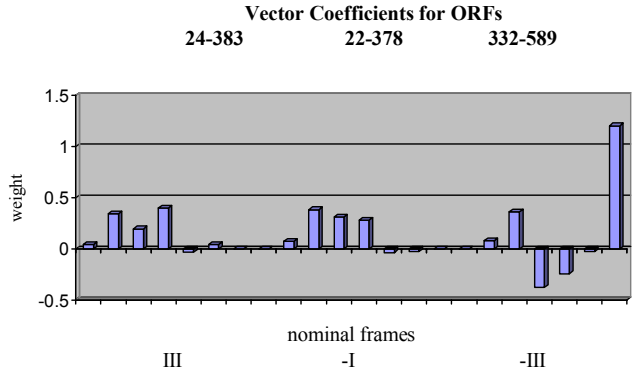
nominal frame	from	to	length (bp)
+III	24	383	360
- I	22	378	356
-III	332	589	258
+III	411	645	236
+ II	389	595	207
-III	2	208	207
- II	45	149	105

(a) The seven ORFs predicted with the NCBI ORF finder for GenBank entry BG056643, representing the EST of an unknown protein. Frames are labelled as in Figure 1. (Note that the NCBI ORF finder uses a different scheme for labelling frames).

(b)



(c)



(b) Graph representing the probabilities of occurrence of the three largest ORFs, in each of their reading frames, calculated using the model.

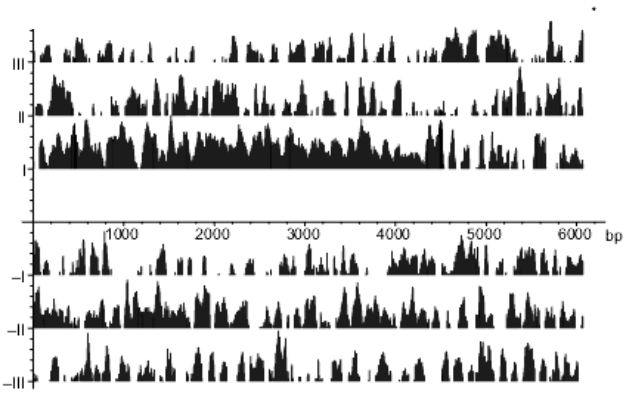
(c) The vector coefficient graphs of the observed probabilities. The six bars in each graph correspond to reading frames 1, 2, 3, -1, -2, -3 (cf., Figure 2). Only the ORF from bp 332-589 (shaded black in part (a) fits the predicted pattern for a piece of coding DNA in its correct frame. Note the large bar in reading frame -3 in the vector coefficient graph for this ORF. (The graph has been normalized for the length of each sequence.)

5.2 Sequence confirmation This analysis is useful for examining the frame of a given cDNA sequence. An example is shown in Figure 4(a), which examines the complete coding sequence of the cystic fibrosis (CFTR) gene (GenBank accession number NM_000492) . The gene is broken into substrings of length 20 trinucleotides, each of which is analyzed using the model. The resulting vector coefficients, c_j, for all the substrings, are plotted as vertical segments on six separate axes. When the coefficients for all the substrings are plotted consecutively, coding DNA will exhibit extended regions with large values of the coefficients in the coding frame.

It is clear that the coding sequence consistently exists in nominal frame I, as expected. One mutant allele of this gene, which leads to the development of cystic fibrosis in homozygotes, has an insertion of two base pairs at position 2560 [5]. When this sequence is analyzed using the model and vector coefficient graphs, the resultant frame shift can clearly be seen (Figure 3 (b); the coding sequence switches from frame I to frame III. This example shows how this algorithm can be used to examine sequences of novel genes, to confirm if they are coding in the same frame throughout, and if not, to identify the approximate positions at which frame shifts have occurred.

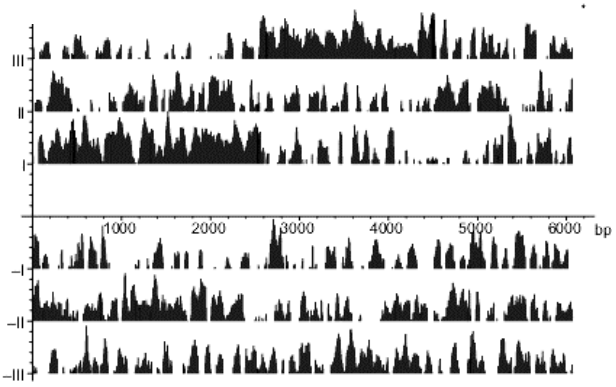
Figure 4 Using the model for sequence confirmation

(a)



(a) Analysis of the complete coding sequence for the human cystic fibrosis (CFTR) gene (coding from bp 133-4575). The height of each vertical bar represents the vector coefficient in a particular frame for a 60 bp substring of the gene. It is clear that the coding sequence exists consistently in frame +I, as expected.

(b)



(b) Analysis of the complete coding sequence for the mutated allele of the human CFTR gene using the model. The two nucleotides inserted in this sequence at bp 2560 produce a frame shift mutation, to frame +III, which is clearly seen.

From the codon frequency database, the frequencies of the stop codons TAA, TAG, and TGA in the coding frame (frame 1) are seen to be 0.0007, 0.0005, and 0.0013, respectively. This means that on average, a string of coding DNA will be $1/(0.0007 + 0.0005 + 0.0013) = 1/(0.0025) = 400$ codons long (i.e., the expected number of codons read before encountering a stop codon is 400). In frame 2, the corresponding frequencies are calculated to be 0.0073, 0.0086, and 0.0278, so for DNA coding in frame 1, but frame shifted to frame 2, the expected

length of a string before encountering a stop codon is 1/(0.0073 + 0.0086 + 0.0278) = 1/(0.0437) ~23 codons. In frame 3, the frequencies are 0.0203, 0.0115, and 0.0255, so the expected length is 1(573) ~17 codons. Similarly, the expected lengths to the first stop codon for frames -1, -2, and -3 are approximately 38, 32, and 19.

It is commonly observed that when a gene displays an insertion or deletion mutation which alters its reading frame, the resulting protein product is severely truncated and thus non-functional. The above description provides a mathematical explanation of this phenomenon. For example, in the CFTR insertion mutation in Figure 4 (b), a stop codon appears in nominal frame I as early as the 12th codon after the insertion mutation, and there are 54 new stop codons before the end of the gene.

5.3 Examining hypothetical genes The model can also be used as a tool to further predict the validity and accuracy of hypothetical genes. While most gene prediction programs used to predict hypothetical genes provide a good indication of the exons in a given gene, the exact boundaries of each exon are often not accurately defined. Joining together these exon sequences could thus lead to frame shift mutations in the predicted gene [6]. In the cases of predicted genes, it is important to determine an accurate sequence and an accurate ORF, so that a prediction of the protein product can be made, and its structure and function can also be hypothesized. In addition, gene prediction programs often provide many results which are false positives; they are not coding sequences at all. A simple method of analyzing results to test for false positives would be very useful. Our model can help to address this problem. A study of 9 hypothetical genes from a region of chromosome 1q22 is shown in Table 1. The results show that five of the sequences clearly fit the model with a high matching score in the appropriate frame, suggesting that they are likely to be coding sequences. Three sequences fit the model poorly and are thus predicted not to be accurate coding sequences. The remainder are inconclusive (i.e., fit the model with 0.50 - 0.75 matching score). These may be parts of genes, sequences with frame shift or other mutations, or they may be non-coding DNA. This model can be employed to narrow down a list of potential candidate genes, to provide an idea of what to focus on first in laboratory experiments. Since this analysis was completed, several of these GenBank entries have been revised or removed (indicating that the predictions were incorrect or false positives). Remarks about these revisions are provided in the comments section of the table, and generally support the predictions of the model. We have developed a Windows software program which is a user-friendly way of applying our mathematical model to predicted genes to test for these false positives. A downloadable version of this program, called FrameView, is available at www.math.toronto.edu/repka/software.

DISCUSSION

Regions of coding DNA should exhibit codon frequencies which conform to the known codon usage tables. The algorithm developed here goes a step beyond simply calculating codon frequencies in a region. The knowledge of codon frequencies in human coding DNA was used to predict trinucleotide frequencies in each of the five noncoding reading frames. A mathematical model was then developed which predicts the probability with which a hypothetical piece of coding DNA can be expected to occur in its coding frame and in each of the other five reading frames. Then, for any given string of DNA, the probability that this string would arise at random in each of the six reading frames of coding DNA is calculated, and these probabilities are compared with the values predicted by the model for regions of coding DNA. The results provide a trinucleotide bias profile of the sequence in all frames.

In earlier work [7], the idea of determining trinucleotide frequencies in non-coding reading frames was introduced. However, since only a limited number of gene sequences were available, the codon frequencies were much less reliable. The calculations often depended on the assumption that genes involved in similar processes should show similar codon frequencies; this might distort the data. By examining an unknown sequence in small (~ 20 trinucleotide) consecutive "windows" in each of the three positive reading frames, they determined, using Bayes' theorem (rather than the more sophisticated model used here), how well each of the consecutive windows corresponded to the expected trinucleotide bias. Their approach was used mainly to search for genes among the vast stretches of non-coding DNA in humans, and was not developed further.

Codon frequency tables for coding DNA have now been much more accurately determined, and are available for many species. This has allowed us to extend the previous work in several ways, including determining accurate trinucleotide frequency tables for non-coding reading frames, and developing a model for analyzing trinucleotides in all six reading frames (i.e., including the three frames read on the reverse complement strand) to produce a signature. Because more information is acquired from a given section of DNA by considering all six reading frames simultaneously, it is possible to obtain meaningful results on shorter strings using this method. While the correct reading frame is expected to yield the highest probability for a piece of coding DNA, the signature of the probabilities from all six reading frames is also significant. It can be used simply to identify coding DNA, but it also makes possible novel practical and theoretical analyses and comparisons (see below). This model is easily adapted for use in any species. The vector analysis used here also makes the data easier to read, providing an immediate indication of agreement with expected results, and also allows rapid visualization of interesting patterns in the data.

Most other algorithms that examine codon bias are not stand alone programs, but are part of larger, more complicated software packages (e.g., gene finding programs), and are based on Hidden Markov Models (HMMs), which typically examine dicodon (i.e., pairs of codons) frequencies in DNA. There are several important differences in our approach. HMMs typically examine the probability in each frame independently, unlike our model, where all six frames are considered simultaneously to determine an overall pattern. Two studies have been performed which use HMMs in all six frames [8.9], but these studies focus on finding specific motifs identifying the start sites for overlapping prokaryotic genes, and do not incorporate the use of codon bias. Compared to HMM methods, our model is computationally much simpler and requires less computer resources. Since HMMs use dicodon frequencies, there are 64 x 64 different dicodon combinations possible. The use of simpler codon frequency tables also makes our model easy to adapt to species other than humans. Our model can thus be used to supplement existing tools. An example of this is the computational identification and analysis of hypothetical genes, and elimination of false positives.

We are currently investigating the application of this model to several other important theoretical and practical issues in biology. Comparisons of codon bias profiles from different species and across organelles (nuclei, mitochondria, chloroplasts) are being used to address evolutionary questions. These codon bias profiles may be able to provide information about the estimated levels of gene expression, and aid in the

identification of various classes or families of genes. Our model of codon bias also inherently considers the related but distinct issue of amino acid bias, that is, the relationship of amino acids usage across different types of genes, organelles, and species, and the model is being further explored to analyze proteins in this way.

Table 1 Using the codon bias model to examine hypothetical genes in the chromosome 1q22 region

table entry #	name of hypothetical gene	GenBank accession #	coding sequence (bp)	nominal frame	matching score for nominal frame	fits model?	comments
1	FLJ23040	XM_043575	125-310	II	0.65	inconclusive	found to be part of entry #2 and removed from GenBank
2	FLJ23040	NM_025174	3-1037	III	0.67	inconclusive	supported by mRNA alignment
3	LOC92299	XM_044075	1125-1460	III	0.38	no	
4	FLJ12671	XM_044082	184-1245	I	0.93	yes	both found to be part of sequence XM_044083 and replaced by this in GenBank; supported by alignment with mRNA and ESTs
5	FLJ12671	XM_044081	90-443	III	0.92	yes	
6	MGC13038	XM_044107	568-756	I	0.55	inconclusive	removed from GenBank
7	LOC200180	XM_010522	1165-1584	I	0.59	inconclusive	supported by alignments with mRNA
8	LOC92306	XM_044124	10-348	I	0.39	no	
9	LOC92307	XM_044125	1762-2265	I	0.06	no	removed from GenBank
10	FLJ20203	NM_032292	1-2508	I	0.87	yes	
11	FLJ20203	XM_043572	564-896	III	0.66	inconclusive	had sequencing errors/frame shift mutations; re-sequenced and found to be same as entry #10; removed from GenBank
12	LOC86036	XM_043567	100-1371	I	0.90	yes	first thought to be a distinct locus, but found to be part of entry #10; removed from GenBank
13	LOC86036	XM_017059	91-1311	I	0.92	yes	first thought to be at same locus as entry #12, but used different start, stop, with frame shift mutations; now found to be unique gene FLJ10875

Table 2 Codon frequencies in coding DNA.

a) Codon frequencies in frame 1 (-1). Parentheses refer to frame -1. The rows indicate the first two nucleotides in frame 1 while the columns give the last. For frame -1, the columns give the first nucleotide and the rows the last two. These are the values reported (Nakamura *et al.,* 2000) as of July 15, 2001.

	***a (t**)*	***c (g**)*	***g (c**)*	***t (a**)*
aa (*tt)*	0.0240	0.0198	0.0326	0.0170
ac (*gt)*	0.0149	0.0193	0.0063	0.0129
ag (*ct)*	0.0115	0.0193	0.0113	0.0120
at (*at)*	0.0072	0.0216	0.0223	0.0158
ca (*tg)*	0.0120	0.0149	0.0345	0.0105
cc (*gg)*	0.0167	0.0200	0.0070	0.0173
cg (*cg)*	0.0063	0.0108	0.0116	0.0046
ct (*ag)*	0.0070	0.0193	0.0397	0.0128
ga (*tc)*	0.0291	0.0261	0.0402	0.0224
gc (*gc)*	0.0159	0.0283	0.0075	0.0185
gg (*cc)*	0.0164	0.0227	0.0164	0.0108
gt (*ac)*	0.0070	0.0146	0.0288	0.0109
ta (*ta)*	0.0007	0.0158	0.0005	0.0121
tc (*ga)*	0.0119	0.0175	0.0045	0.0148
tg (*ca)*	0.0013	0.0123	0.0129	0.0100
tt (*aa)*	0.0073	0.0205	0.0125	0.0170

b) Calculated trinucleotide frequencies in frame 2 (-3):

	a (t)	**c (g**)	**g (c**)	**t (a**)
aa (*tt)*	0.0176	0.0161	0.0208	0.0113
ac (*gt)*	0.0205	0.0188	0.0242	0.0131
ag (*ct)*	0.0289	0.0264	0.0340	0.0185
at (*at)*	0.0166	0.0152	0.0196	0.0106
ca (*tg)*	0.0159	0.0146	0.0187	0.0102
cc (*gg)*	0.0228	0.0208	0.0269	0.0146
cg (*cg)*	0.0068	0.0062	0.0080	0.0043
ct (*ag)*	0.0170	0.0156	0.0200	0.0109
ga (*tc)*	0.0095	0.0087	0.0112	0.0061
gc (*gc)*	0.0174	0.0160	0.0205	0.0112
gg (*cc)*	0.0140	0.0128	0.0165	0.0090
gt (*ac)*	0.0100	0.0092	0.0118	0.0064
ta (*ta)*	0.0076	0.0070	0.0090	0.0049
tc (*ga)*	0.0204	0.0186	0.0240	0.0130
tg (*ca)*	0.0277	0.0253	0.0326	0.0177
tt (*aa)*	0.0151	0.0138	0.0178	0.0097

c) Calculated trinucleotide frequencies in frame 3 (-2):

	a (t)	**c (g**)	**g (c**)	**t (a**)
aa (*tt)*	0.0177	0.0101	0.0102	0.0126
ac (*gt)*	0.0136	0.0115	0.0063	0.0149
ag (*ct)*	0.0223	0.0133	0.0125	0.0116
at (*at)*	0.0055	0.0092	0.0069	0.0108
ca (*tg)*	0.0283	0.0162	0.0164	0.0203
cc (*gg)*	0.0218	0.0185	0.0101	0.0239
cg (*cg)*	0.0358	0.0213	0.0201	0.0186
ct (*ag)*	0.0088	0.0148	0.0111	0.0174
ga (*tc)*	0.0270	0.0154	0.0156	0.0193
gc (*gc)*	0.0208	0.0176	0.0096	0.0228
gg (*cc)*	0.0341	0.0203	0.0192	0.0177
gt (*ac)*	0.0084	0.0141	0.0105	0.0166
ta (*ta)*	0.0205	0.0117	0.0118	0.0147
tc (*ga)*	0.0157	0.0134	0.0073	0.0173
tg (*ca)*	0.0258	0.0154	0.0145	0.0134
tt (*aa)*	0.0064	0.0107	0.0080	0.0125

REFERENCES

[1] S. Zhang and G. Zubay, "The peculiar nature of codon usage in primates", **Genetic Engineering**,13, 73-113, 1991.

[2] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for The year 2000", **Nucleic Acids Research,** 28, 292, 2000.

[3] Rainey, S. and Repka, J. "Codon bias as a quantitative tool for DNA sequence analysis". **Journal of Systemics, Cybernetics and Informatics**, Proceedings July 2005

[4] J.W. Fickett, "Finding genes by computer: the state of the art", **Trends in Genetics,** 12, 316-320, 1996.

[5] M.B. White, J. Amos, J.M. Hsu, B. Gerrard, P. Finn, and M. Dean, "A frame-shift mutation in the cystic fibrosis gene", **Nature,** 344, 665-667, 1990.

[6] E.E. Snyder and G.D. Stormo, "Identification of protein coding regions in genomic DNA", **Journal of Molecular Biology,** 248, 1-18. 1995.

[7] R. Staden and A.D. McLachlan, "Codon preference and its use in identifying protein coding regions in long DNA sequences", **Nucleic Acids Research,** 10, 141-156, 1982.

[8] A.M. Shmatkov, A.A. Melikyan, F.L. Chernousko, and M. Borodovsky, "Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes", **Bioinformatics,** 15, 874-886, 1999.

[9] A.V. Lukashin and M. Borodovsky, "GeneMark.hmm: new solutions for gene finding", **Nucleic Acids Research, 26,** 1107-1115, 1998.