Prediction of probable genes by Fourier analysis of genomic sequences

Shrish Tiwari¹, S. Ramachandran², Alok Bhattacharyya³, Sudha Bhattacharya² and Ramakrishna Ramaswamy¹,⁴

Abstract

Motivation: The major signal in coding regions of genomic sequences is a three-base periodicity. Our aim is to use Fourier techniques to analyse this periodicity, and thereby to develop a tool to recognize coding regions in genomic DNA.

Result: The three-base periodicity in the nucleotide arrangement is evidenced as a sharp peak at frequency \( f = 1/3 \) in the Fourier (or power) spectrum. From extensive spectral analysis of DNA sequences of total length over 5.5 million base pairs from a wide variety or organisms (including the human genome), and by separately examining coding and non-coding sequences, we find that the relative height of the peak at \( f = 1/3 \) in the Fourier spectrum is a good discriminator of coding potential. This feature is utilized by us to detect probable coding regions in DNA sequences, by examining the local signal-to-noise ratio of the peak within a sliding window. While the overall accuracy is comparable to that of other techniques currently in use, the measure that is presently proposed is independent of training sets or existing database information, and can thus find general application.

Availability: A computer program GeneScan which locates coding open reading frames and exonic regions in genomic sequences has been developed, and is available on request. Contact: E-mail: rama@jnuniv.ernet.in.

Introduction

The gene identification problem (Fickett, 1996), namely the identification of protein-coding genes in DNA sequences through computational means, is of great current importance. The worldwide initiative on genome sequencing has necessitated the development of new approaches to assess rapidly the potential of a given nucleotide sequence in a functional context. Genome projects have given rise to an exponentially growing amount of genetic information, much of which is novel: even in a simple eukaryote like Saccharomyces cerevisiae, less than half the number of potential gene sequences were known prior to the recently completed Yeast Genome Project.

A number of methods have been proposed for gene detection, based on distinctive features of protein-coding sequences. These have recently been comprehensively reviewed (Fickett and Tung, 1992; Fickett, 1996). The different methods are based on a variety of contrasting characteristics of protein-coding DNA sequences and DNA sequences that do not encode proteins. These methods employ, for example, differences in the patterns of codon usage (Staden and McClellan, 1982) or oligonucleotide frequencies (Shulman et al., 1981, Fickett, 1982; Borodovsky et al., 1986, 1994), or train neural nets (Lapedes et al., 1990; Uberbacher and Mural, 1991; Farber et al., 1992; Xu et al., 1994; Snyder and Stormo, 1995) to recognize the distinctive features of the two sets. Other techniques use linguistic methods (Dong and Searls, 1994; Mantegna et al., 1994) and correlation functions (Li, 1992; Peng et al., 1993; Ossadnik et al., 1994; Buldyrev et al., 1995). At the same time, comprehensive evaluations of the various methods suggest that they cannot be expected to work equally well for all genes (Burset and Guigo, 1996), and constant refinement is needed to evolve better methodologies. There is also a need (Fickett, 1996) for new methods of gene prediction which utilize features of gene structure that have not so far been incorporated in programs already available.

In this paper, we investigate a Fourier technique based on a distinctive feature of protein-coding regions of DNA sequences, namely the existence of short-range correlations in the nucleotide arrangement. The most prominent of these is a 1/3 periodicity, which has been shown to be present in coding sequences (Fickett, 1982). The signature of this (and indeed any other) periodicity can be seen most directly through the Fourier analysis (Tsonis et al., 1991; Voss, 1992) as a spectral peak. In the present work, we analyse genomic sequences from different organisms, and verify that such periodicity is universal for protein-coding sequences and is absent in genomic sequences which do not code for proteins. The quantitative measure that we focus on is the relative strength of this periodicity, which we then use in order to develop a simple technique to predict genes (with and without introns) in unknown genomic sequences of any organism. The present method, like other Fourier-based methods, offers some advantages, namely that it is quite easy to apply and requires no prior knowledge of the sequence to be analysed.

The origin of the period-3 signal in protein-coding sequences derives from the triplet nature of the codon. This

¹School of Physical Sciences.
²School of Environmental Sciences and
³School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India
⁴To whom correspondence should be addressed
Table I. Summary of genomic sequences studied

<table>
<thead>
<tr>
<th>Group</th>
<th>Species</th>
<th>G + C (%)</th>
<th>ORFs(^a)</th>
<th>Genes(^b)</th>
<th>ORFs with (P \geq 4)</th>
<th>Genes with (P \geq 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fungus</td>
<td><em>S. cerevisiae</em> III</td>
<td>38.6</td>
<td>216</td>
<td>54</td>
<td>198</td>
<td>51</td>
</tr>
<tr>
<td>Fungus</td>
<td><em>S. cerevisiae</em> VIII</td>
<td>38.2</td>
<td>267</td>
<td>140</td>
<td>255</td>
<td>139</td>
</tr>
<tr>
<td>Insect</td>
<td><em>A. californica</em></td>
<td>40.7</td>
<td>154</td>
<td>51</td>
<td>137</td>
<td>49</td>
</tr>
<tr>
<td>Virus</td>
<td><em>E. histolytica</em></td>
<td>34.2</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Bacteria</td>
<td><em>A. vinelandii</em></td>
<td>62.4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Bacteria</td>
<td><em>H. influenzae</em></td>
<td>38.2</td>
<td>1727</td>
<td>933</td>
<td>1667</td>
<td>927</td>
</tr>
<tr>
<td>Nematode</td>
<td><em>C. elegans</em></td>
<td>35.6</td>
<td>–</td>
<td>146</td>
<td>–</td>
<td>146</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>35.1</td>
<td>24</td>
<td>–</td>
<td>24</td>
<td>–</td>
</tr>
<tr>
<td>Various</td>
<td>Globins(^c)</td>
<td>49.6</td>
<td>–</td>
<td>15</td>
<td>–</td>
<td>15</td>
</tr>
<tr>
<td>Various</td>
<td>Actins(^d)</td>
<td>36.8</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

\(^a\)Total number of ORFs as reported in the literature.
\(^b\)ORFs positively identified as coding for proteins through homology, detection of the corresponding c-DNA.
\(^c\)The G + C content quoted is the average over genes only. Data from Sehgal *et al.* (1994) as well as GenBank. The G + C content is the average over the sequences analysed.
\(^d\)H.K. Das, private communication.
\(^e\)The human sequences used in this study are those used by Uberbacher and Mural (1991) for GRAIL.

fact alone is, however, insufficient to explain why coding regions exclusively have the signal, while non-coding regions overwhelmingly do not, and the reasons for this distinction lie in the unequal usage of codons (codon bias) in coding regions (Tsonis *et al.*, 1991), as well as in the biased usage of nucleotide triples in genomic DNA (the triplet bias). The latter bias comes in part from the unequal usage of the amino acids in naturally occurring proteins, and is universally present. The former bias arises from the unequal usage of the codons corresponding to a given amino acid, and is specific to a given organism. In order to explore the role of the two types of compositional bias in generating the period-3 signal, we have performed a variety of numerical experiments. Our results indicate that while codon bias does play a role, it is, however, not the primary reason for the periodicity.

**Algorithm**

The Fourier analysis described below has been performed on nucleotide sequences, of total length over 5.5 Mbases. These sequences (listed in Table I) include the complete sequences of yeast (*S. cerevisiae*) chromosomes III (Oliver *et al.*, 1992).

![Fig. 1](https://academic.oup.com/bioinformatics/article-abstract/13/3/263/423173/264)
and VIII (Johnson et al., 1994); the genome of Autographa californica nuclear polyhedrosis virus (Ac-NPV) (Ayres et al., 1994); 2.2 Mb of contiguous sequences of Caenorhabditis elegans chromosome III (Wilson et al., 1994); several cDNA and plasmid sequences of a protozoan parasite Entamoeba histolytica (Sehgal et al., 1994; as well as GenBank); the genome of Haemophilus influenzae (Fleischmann et al., 1995); a few genes from the bacterium Azotobacter vinelandii (H.K.Das, private communication), and human genomic sequences (Bilofsky and Burks, 1988).

A sequence of N nucleotides may be formally viewed as a symbol string, \( \{x_j\}, \) where \( x_j \) is one of the four symbols A, T, G and C, and denotes the occurrence of that particular nucleotide in position \( j. \) In order to define the Fourier spectrum for a genomic sequence, we adopt the following procedure. One can define a binary indicator function or projection operator \( \Lambda (\alpha) \) which selects the elements of the sequence that are equal to the symbol \( \alpha, \) namely \( S_\alpha (x_j) = 1 \) if \( x_j = \alpha \) and 0 otherwise. Using the operators \( \Lambda_A, \Lambda_T, \Lambda_G \) and \( \Lambda_C \) successively on a DNA sequence yields four binary sequences, as illustrated below:

Sequence: G G A T A T C A C T T T T A G A G

Apply \( \Lambda_A: \) 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0

Apply \( \Lambda_T: \) 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0 0

Apply \( \Lambda_G: \) 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1

Apply \( \Lambda_C: \) 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0

Thus, any DNA sequence can be converted to four binary sequences, which can then be Fourier analysed in the normal manner, to examine correlations between the symbols. The total Fourier spectrum of the DNA sequence is (Silverman and Linsker, 1986; Li et al., 1994) the sum of these individual spectra, namely:

\[
S(f) = \sum_\alpha S_\alpha(f) = \sum_\alpha \frac{1}{N^2} \left| \sum_{j=1}^N \Lambda_\alpha (x_j) \exp 2 \pi j f \right|^2
\]

(1)

where the discrete frequency \( f = k/N, \) with \( k = 1, 2, \ldots, N/2. \) \( S_\alpha(f) \) is the partial spectrum corresponding to the symbol \( \alpha = A, T, G \) or \( C. \) The average of the total spectrum, \( \bar{S}, \) can be calculated (see, for example, Chechetkin and Turygin, 1995) from the frequency of occurrence, \( \rho_\alpha \) of each symbol (\( \alpha = A, T, G, C \) ) as:

\[
\bar{S} = \frac{2}{N} \sum_{k=1}^{N/2} S(k/N) = \frac{1}{N} \left( 1 + \frac{1}{N} - \sum_\alpha \rho_\alpha^2 \right)
\]

(2)

For protein-coding sequences from a variety of organisms, the Fourier spectrum [equation (1)] reveals the characteristic periodicity of three as a distinct peak at frequency \( f = 1/3 \) (a typical pattern is shown in Figure 1a). No such 'peak' above the noise level is apparent for non-protein coding sequences such as rRNA, intergenic spacers and introns, which have a flat Fourier spectrum devoid of any periodicity (see Figure 1b). In order to contrast the two types of spectra, we focus on the signal-to-noise ratio of the peak at \( f = 1/3, \) namely:

\[
P = S(1/3)/\bar{S}
\]

(3)

Our survey of a large number of coding and non-coding sequences from a variety of organisms is summarized in Table I, which gives details of the systems studied, and Figure 2, which shows cumulative distributions of the signal-to-noise ratio for coding and non-coding sequences. The solid curve in Figure 2 shows the fraction of coding sequences with \( P < 4 \) less than the abscissa, and the dashed curve shows the fraction of non-coding sequences with \( P > 4 \) greater than the abscissa. The two distributions can be clearly seen to have only a small area of overlap: this suggests, as is made explicit below, that a coding measure can be devised from this observation.

We use the value \( P = 4 \) as a discriminator between coding and non-coding sequences. From the data presented in Table I and Figure 2, it is evident that the bulk of coding sequences (~95%) from all organisms tested so far satisfy this criterion. Similarly, almost 90% of the non-coding regions have \( P > 4; \) these sequences were taken from the intergenic regions of the several organisms used in this study, including S.cerevisiae chromosomes III and VIII, baculovirus (A.californica), H.influenzae, E.histolytica (a total of >1.0 Mbase). The bars in Figure 2 indicate the variation in values of \( P \) from organism to organism; there appears to be no particular systematics, since the signal-to-noise ratio depends strongly on the length of the sequence.

One set of counterexamples that we have seen to the above coding sequences with \( P < 4. \) The few such cases are instructive: for instance, the Mata proteins alone in S.cerevisiae chromosome III do not show any peak. Whether this fact is related to unusual amino acid organization of these proteins is currently being investigated. The other set, namely non-coding sequences with \( P > 4, \) are slightly more common, but these can often be easily recognized as non-coding from the existence of several other periodic features (the whole spectrum appears more grassy, in contrast to coding sequences where the \( f = 1/3 \) peak is invariably the only prominent spectral feature). From our analysis so far, the non-coding sequences with \( P > 4 \) do not appear to belong to any specific sequence category.

In order to utilize this measure to predict potential protein-coding sequences, we first note that the spectral approach can be applied even to fairly short gene sequences (even of the order of a few hundred bases). A genomic sequence of unknown functionality can be analysed for putative protein-coding properties by the following procedure. An \( M \) nucleotide sequence window of the complete sequence is analysed according to equation (1), and the existence of a
peak at $f = 1/3$ can be used to identify whether this subsequence forms part of a coding or non-coding region. For each window, the local signal to noise ratio, $P_M(j)$, is computed according to equation (3) ($j$ being the position of the centre of the window of length $M$). Note that this involves the calculation of a single spectral line, and not the total spectrum.

By sliding this window along the sequence, we generate a graph of $P_M(j)$ versus $j$ which makes it possible essentially to 'read off' the probable coding regions: these are those portions of the sequence with $P_M > 4$. To identify the exact end points of the coding region, the sequence is scanned to a distance of $M$ nucleotides up- and downstream of the above identified region so as to locate the initiation and termination codons, respectively, in any of the six possible reading frames. Having obtained these endpoints, we generate the total spectrum [equation (1)] of the putative gene in order to verify that the spectral feature of $f = 1/3$ is distinctive and characteristic of a coding region. This double check is useful in reducing the number of false positives.

The window length, $M$, and the discrimination value need to be chosen in any implementation of the method. In our study, we have taken the window length as $M = 351$ for the analysis of genomic sequences of yeast and insect virus, since there are very few intron-containing genes in these sequences, and open reading frames (ORFs) of length less than 300 bp are not frequently encountered. A window length in the range of 250–400 gives similar results, although the number of false negatives in the sequences studied was least for 351. Windows of length less than 250 have increased noise and somewhat poorer statistics, while those greater than 400 tend to miss the ORFs due to numerous overlaps. For higher organisms, where we expect introns, shorter windows ($M \sim 150$) were used. As regards the discrimination value, an alternate way of deciding the optimal value of $P$ to use as a discriminator would be to take the minimum in the sum of the two curves in Figure 2. Our results indicate that using a value of $P_M = 5$ would not give drastically different results. Lower values of the threshold increase the number of false positives—we get more and more overlapping windows, and subsequent ORF analysis tends to pick up all ORFs, whether coding or not.

**Implementation**

In this section, the technique described above is applied to a variety of genomic sequences from several different organisms. The results of our analysis of yeast chromosomes III and
Gene prediction by Fourier analysis

Table II. Summary of results for S.cerevisiae chromosomes III and VIII, and H.influenzae

<table>
<thead>
<tr>
<th></th>
<th>Chromosome III</th>
<th>Chromosome VIII</th>
<th>H.influenzae</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORFs</td>
<td>216</td>
<td>267</td>
<td>1727</td>
</tr>
<tr>
<td>ORFs detected</td>
<td>187</td>
<td>226</td>
<td>1499</td>
</tr>
<tr>
<td>False positives detected</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.87</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>Genes reported</td>
<td>54</td>
<td>140</td>
<td>933</td>
</tr>
<tr>
<td>Genes detected</td>
<td>44</td>
<td>123</td>
<td>867</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.81</td>
<td>0.88</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table IIIA. Summary of results for human and C.elegans genomic sequences

<table>
<thead>
<tr>
<th></th>
<th>C.elegans</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes reported</td>
<td>146</td>
<td>24</td>
</tr>
<tr>
<td>Genes detected</td>
<td>146</td>
<td>24</td>
</tr>
<tr>
<td>Exons reported</td>
<td>982</td>
<td>141</td>
</tr>
<tr>
<td>Exons detected</td>
<td>837</td>
<td>119</td>
</tr>
<tr>
<td>Exons &gt; 100bp reported</td>
<td>844</td>
<td>93</td>
</tr>
<tr>
<td>Exons &gt; 100bp detected</td>
<td>764</td>
<td>86</td>
</tr>
</tbody>
</table>

Table IIIB. Summary of results for exon detection in ALLSEQ (Burset and Guigo, 1996)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Approximate correlation</th>
<th>Missing exons</th>
<th>Wrong exons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.66</td>
<td>0.60</td>
<td>0.53</td>
<td>0.31</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Indeed confirmatory, the number of false positives is drastically reduced; in this instance, the number is actually zero and thus the specificity [specificity = number detected/(number detected + false positives)] is exactly 1.00. A few genes, for example the mating type, do not show this periodicity, and thus will not be identifiable through our analysis. Similar observations hold for the analysis of the H.influenzae genome.

Our method has been applied to genes that contain introns as follows. We first applied our technique to the 2.2 Mbase C.elegans and several human genomic sequences, as well as globin sequences. In this analysis, we shortened the window size to M = 150, and this sets the limit on the shortest exonic regions that can be predicted with confidence to ~ 100 bp. A representative analysis using goat β globin is shown in Figure 3b. The three exonic regions and the intermediate introns can be clearly distinguished, and precise boundaries of the exons can be demarcated by using the canonical splice acceptor and donor sites.

A summary of our results is given in Table IIIA. All 24 genes reported to be present in the human sequences and all 146 of those in C.elegans could be correctly identified (i.e. at least one exon in each of the genes was identified). Within the genes, we identified 119 out of 141 exons in human sequences, and 837 out of 982 exons in C.elegans. However, when exons shorter than 100 bp were excluded from the analysis, 86 out of 93 such exons could be identified in the former set and 764 out of 844 in the latter, signifying a sensitivity of over 0.9. Our data are comparable to predictions of other methods (Uberbacher and Mural, 1991; Snyder and Stormo, 1995) that locate exonic regions in genomic sequences.

A more stringent test of gene structure prediction is possible, and to evaluate the efficiency of our method vis-à-vis established methods currently in use, we have applied our technique to the set of sequences (ALLSEQ) recently
Fig. 4. The persistence of the spectral signature of myosin from *E. histolytica* upon artificially altering the nucleotide sequence. (a) The Fourier transform of the myosin DNA sequence. (b) The Fourier transform with the codon bias removed, by translating the sequence to the protein, and regenerating a nucleotide sequence with the codons corresponding to a given amino acid used with equal probabilities. (c) The Fourier transform with the genetic code scrambled, namely by translating the sequence to the protein and then assigning an arbitrary genetic code.

compiled by Burset and Guigó (1996), to benchmark a number of different gene-structure prediction programs. We chose a random subset of 75 sequences and obtained the results shown in Table IIIB, which indicate that the sensitivity, specificity and approximate correlation (Burset and Guigó, 1996) of our method GeneScan is comparable to (if somewhat poorer than) programs such as GeneParser2 (0.65, 0.78, 0.65), GenLang (0.70, 0.73, 0.65), GRAIL II (0.70, 0.83, 0.74) and SORFIND (0.68, 0.83, 0.70). However, this comparison may not be entirely appropriate since the dataset ALLSEQ is designed for complete gene recognition and our method is directed toward coding region recognition; we feel that it will be possible to improve our algorithm greatly with further refinements or by incorporation in more sophisticated programs.

**Discussion**

We now briefly address the question of the origin of the spectral signature peak at $f = 1/3$, which is obviously related
to the triplet nature of the genetic code. Beyond this, however, the reasons for the distinction between coding and non-coding genomic sequences are less clear. From our present analysis of coding sequences with widely differing base composition, the periodicity appears not to be a consequence of codon bias. In numerical experiments (see Figure 4) when this bias was removed—by using all codons corresponding to a given amino acid with equal frequency—we observed that the resulting genomic sequences continued to show a sharp peak at \( f = 1/3 \), although with changed \( P \). The peak remained prominent even when the nucleotide sequence was changed completely by assigning arbitrary codons to a given amino acid, or when the four-symbol sequence was mapped into a two-symbol (purine–pyrimidine) sequence. These experiments suggest that the periodicity may be a consequence of the amino acid sequence in naturally occurring proteins (Zhurkin, 1981) which manifests itself as a bias in the use of certain triplets in the coding regions of genomic DNA.

The gene-detection methodology presented here, GeneScan, while offering a level of predictive accuracy which is comparable to other methods currently in use, has some distinct advantages over and above the ease in its implementation. Being based on a universal property of coding sequences, it is independent of a training set of genes from which priors can be estimated. Unlike neural net-based methods (Uberbacher et al., 1996), the method is relatively independent of variations in base composition.

An important feature of our method is that it is robust with respect to sequencing errors resulting in frameshift mutations. We have performed several tests by introducing sequencing errors of up to 1%, when the overall Fourier spectrum changes very marginally. When the number of exons is large, however, errors of up to 1%, when the overall Fourier spectrum changes very marginally.

### Acknowledgements

We have been supported by grants from the Department of BioTechnology to A.B., and Department of Science and Technology and the CSIR, India, to R.R. Some of the sequences analysed were provided by H.K.Das and S E.Hasnain. We thank J.Subbarao, C Chakravarty, D.Sears and R.Guiogó for helpful comments, and, particularly, Mark Borodovsky for generous advice. This manuscript was completed during the 1995 workshop 'Identifying Features in Biological Sequences' held at the Aspen Center of Physics. R.R. would like to thank Wendent Li for helpful suggestions and participants at the workshop for discussions.

### References


S.Tiwari et al.


Wilson, R. et al. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans.* *Nature,* 368, 32–38.


Received on September 15, 1996; accepted on December 31, 1996