CABIOS

Vol. 13 no. 3 1997
Pages 263–270

Prediction of probable genes by Fourier analysis of genomic sequences

Shrish Tiwari[1], S.Ramachandran[2], Alok Bhattacharya[3], Sudha Bhattacharya[2] and Ramakrishna Ramaswamy[1,4]

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 McLachlan, 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 Guigó, 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

[1] *School of Physical Sciences,*
[2] *School of Environmental Sciences and*
[3] *School of Life Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*
[4] *To whom correspondence should be addressed*

Table I. Summary of genomic sequences studied

Group	Species	G + C (%)	ORFs[a]	Genes[b]	ORFs with $P \geq 4$	Genes with $P \geq 4$
Fungus	*S.cerevisiae* III	38.6	216	54	198	51
Fungus	*S.cerevisiae* VIII	38.2	267	140	255	139
Insect virus	*A.californica*	40.7	154	51	137	49
Protozoa	*E.histolytica*[c]	34.2	26	26	26	26
Bacteria	*A.vinelandii*[d]	62.4	6	6	6	6
Bacteria	*H.influenzae*	38.2	1727	933	1667	927
Nematode	*C.elegans*	35.6	–	146	–	146
Mammal	Human[e]	51.2	–	24	–	24
Various	Globins[f]	49.6	–	15	–	15
Various	Actins[g]	36.8	15	15	15	15

[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.
[f] The GenBank locus names of the sequences studied are: ACAACTI, CELACI, DROACT2A, BOVACT1, BOVACT2, BOVACT2, RABACMA1, RABACMA2, CHKACACA, QULACASK, SLMACT15, SLMACT21, SOYACT1G, MUSACACM, MUSACASA, HUMACBPA1. The G + C content quoted is the average over all the actins from different species.
[g] The GenBank accession numbers of sequences studied are: J00153, J00182, J00413, J03082, K01714, K03256, M17601, M17602, M17909, M61740, V00491, V00513, X00371, X00372, X00373, X03248 and X04862. The G + C content quoted is an average over all the globin sequences from different species.

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)

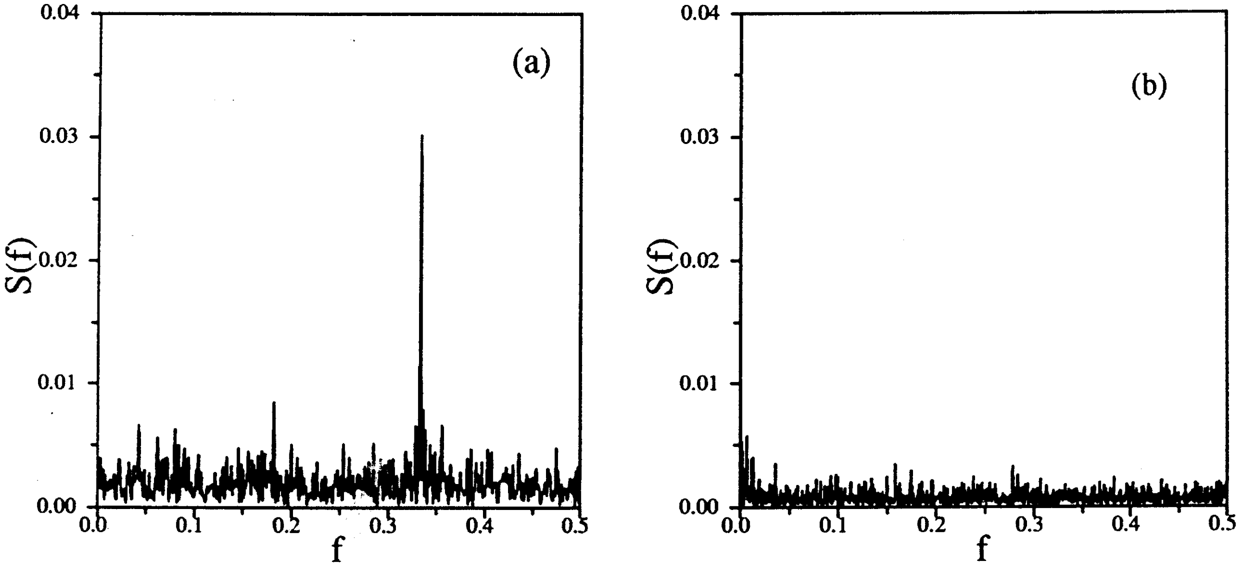

Fig. 1. Typical Fourier spectra for (**a**) a coding stretch of DNA and (**b**) a non-coding stretch from *S.cerevisiae* chromosome III.

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, $\{x_j, j = 1, 2, \ldots, N\}$, 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 (Voss, 1992; see Figure 1 of this paper for a graphical illustration of the use of the projection operator) U_α which selects the elements of the sequence that are equal to the symbol α, namely $U_\alpha(x_j) = 1$ if x_j is α and 0 otherwise. Using the operators U_A, U_T, U_G and U_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	A	G	A	G
Apply U_A	0	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0
Apply U_T	0	0	0	1	0	1	0	0	0	1	1	1	0	0	0	0
Apply U_G	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Apply U_C	0	0	0	0	0	0	1	0	1	0	0	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:

$$S(f) \equiv \sum_{\alpha} S_\alpha(f) = \sum_{\alpha} \frac{1}{N_2} \left| \sum_{j=1}^{N} U_\alpha(x_j) \exp 2\pi i f j \right|^2 \qquad (1)$$

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, $\overline{S}$, can be calculated (see, for example, Chechetkin and Turygin, 1995) from the frequency of occurrence, ρ_α of each symbol ($\alpha = A, T, G, C$) as:

$$\overline{S} \equiv \frac{2}{N} \sum_{k=1}^{N/2} S(k/N) = \frac{1}{N} \left(1 + \frac{1}{N} - \sum_{\alpha} \rho_\alpha^2 \right) \qquad (2)$$

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 peiodicity (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:

$$P = S(1/3)/\overline{S} \qquad (3)$$

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 less than the abscissa, and the dashed curve shows the fraction of non-coding sequences with P 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 are coding sequences with $P < 4$. The few such cases are instructive: for instance, the Matα 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

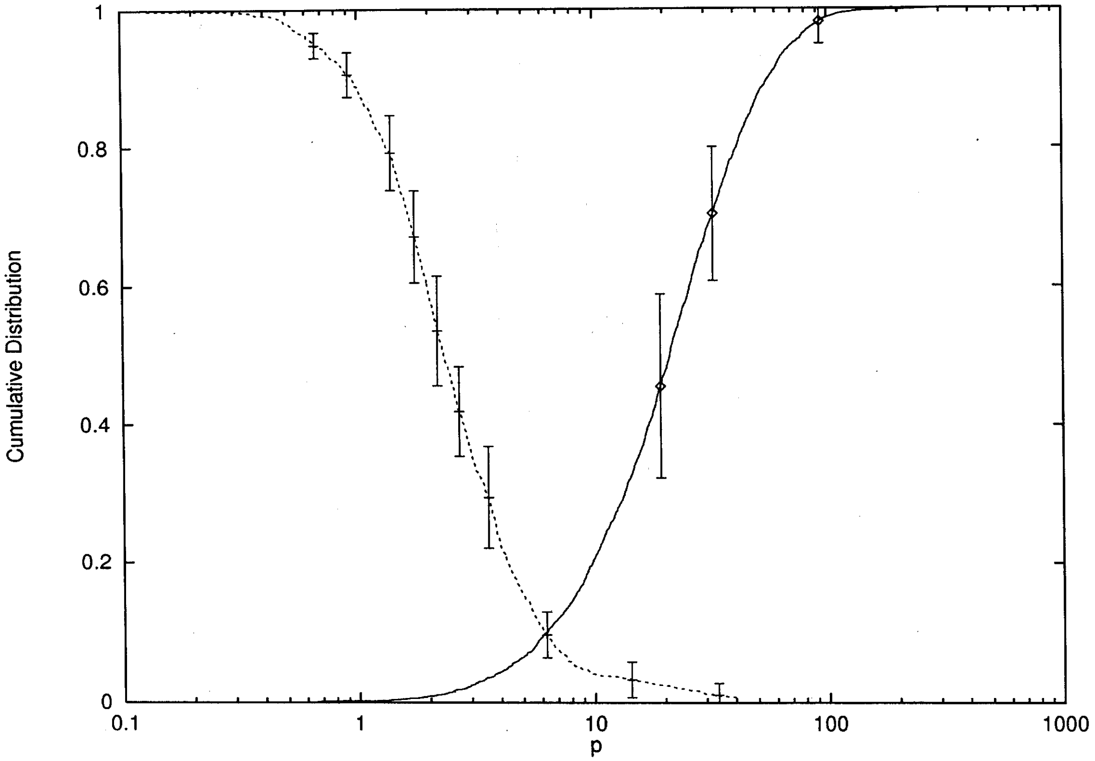

Fig. 2. Cumulative distributions of local peak-to-noise ratio at frequency $f = 1/3$ for the sequences studied. The solid curve (right) is the fraction of coding sequences with P less than the abscissa, while the dashed curve (left) is the fraction of non-coding sequences with P greater than the abscissa. Thus, almost 95% of all coding sequences have $P \geq 4$, while nearly 90% of all non-coding sequences will have $P \leq 4$. The indicated bars show the variation in the distribution among the different organisms studied.

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

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

	Chromosome III	Chromosome VIII	*H.influenzae*
ORFs	216	267	1727
ORFs detected[1]	187	226	1499
False positives detected	0	0	0
Specificity	1.0	1.0	1.0
Sensitivity	0.87	0.85	0.87
Genes reported	54	140	933
Genes detected	44	123	867
Sensitivity	0.81	0.88	0.93

VIII, and the genome of *H.influenzae* are given in Table II. We have used a scanning window of length $M = 351$. Of the 483 probable genes (ORFs) reported in *S.cerevisiae* chromosomes III and VIII (Oliver *et al*., 1992; Johnston *et al*., 1994), our method locates 413 of these exactly. Of the 194 identified genes (i.e. through homology search or detection of corresponding c-DNA), the present method locates 167, giving a sensitivity index (sensitivity = number detected/number reported) of 0.86 at the gene level. By the procedure described in the previous section, namely first identifying a probable gene and then verifying that the Fourier spectrum is 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.

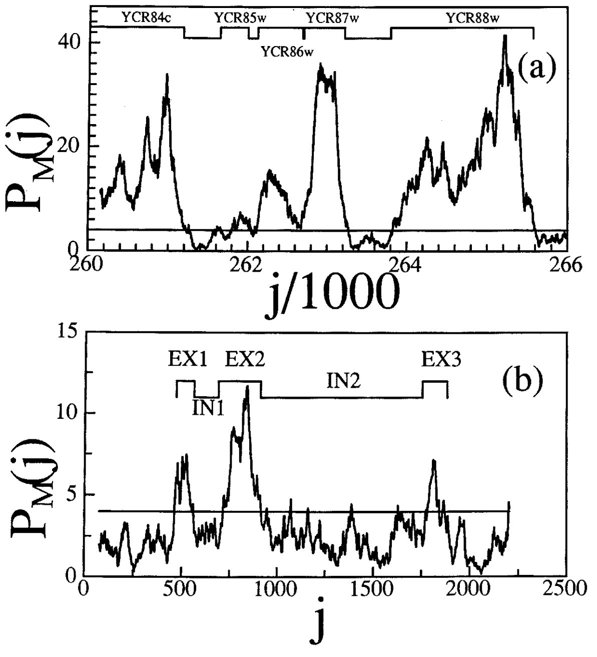

Fig. 3. Window analysis of the local signal-to-noise ratio, $P_M(j)$ *for (***a***)* a portion of the 315 000 nucleotide chromosome III of *S.cerevisiae* from location 260 000 to 265 000. The identified ORFs, from Oliver *et al.* (1992), are indicated at the top of the figure. The length of the window was 351 bases and j varies in steps of three. The baseline is set at $P_M(j) = 4$. (**b**) Window analysis for the β globin of goat, of length 2278 bp—the exons correspond to nucleotides 471–562, 689–911 and 1754–1882; the second intron (911–1754) is much longer than the first intron (562–689). These are indicated at the top of the figure. We used a window of length 150 bases and the baseline is also set at $P_M(j) = 4$.

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

	C.elegans	Human
Genes reported	146	24
Genes detected	146	24
Exons reported	982	141
Exons detected	837	119
Exons > 100 bp reported	844	93
Exons > 100 bp detected	764	86

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

Sensitivity	Specificity	Approximate correlation	Missing exons	Wrong exons
0.66	0.60	0.53	0.31	0.35

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

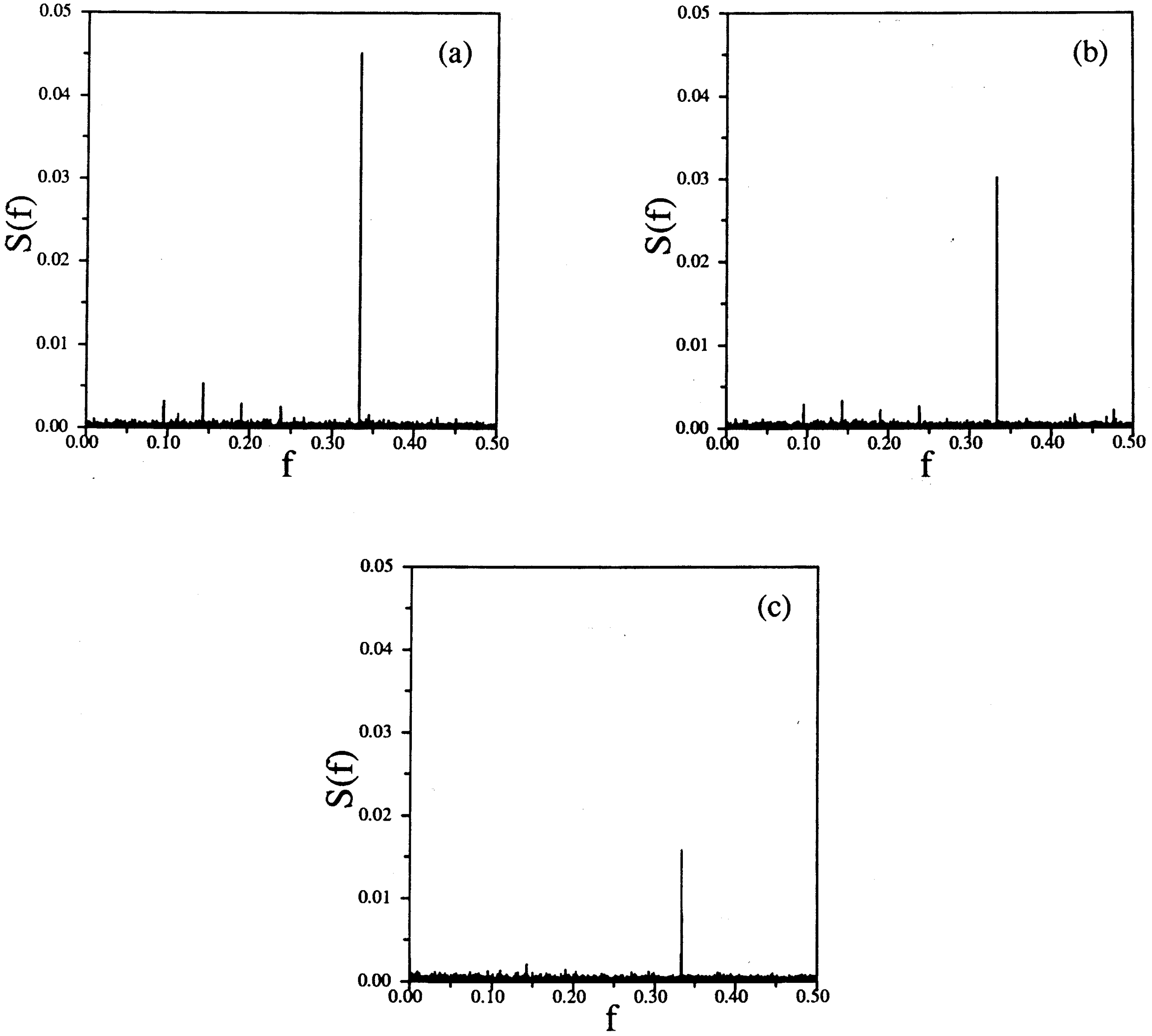

Fig. 4. The persistence of the spectral signature of myosin from *E.histolytica* upon artificially altering the nucleotide sequence. (**a**) The Fourier transform of the myosin DNA sequence. (**b**) The Fourier transform with the codon bias removed, by translating the sequence to the protein, and regenerating 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 and Mural, 1991; Snyder and Stormo, 1995), no *a priori* knowledge of the nature and the character of the sequences that constitute the gene in a specific organism is required. Unlike methods based on correlation exponents (Ossadnik *et al*., 1994), we can work with small genomic sequences (some sequences in ALLSEQ are 450 bases or so in length). In contrast to methods based on codon usage patterns (Borodovsky *et al*., 1986), 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, and there are particularly short exons, though, sequencing and frameshift errors such as insertions and deletions can vitiate the technique as well. Other limitations that should be pointed out are that the boundaries of the exonic regions through Fourier analysis alone (i.e. without searching for conjugate splicing sites) could be fixed only approximately, to within 54 bases on average. Our method can profit by incorporating other, complementary, techniques to locate the precise exon/intron boundaries. Furthermore, the three-base periodicity appears to be lacking in a small percentage of genes, e.g. the genes of the mating-type locus of *S.cerevisiae* and amoebapore of *E.histolytica* among others, which makes them invisible to the spectral analysis. (Exploration of the reasons for the lack of periodicity is currently under way. Preliminary analysis shows that some of these genes have highly restricted amino acid usage.) Given the fact that we do not yet know all that goes into the making of a gene and how a typical gene evolved and, further, that all genes may not necessarily obey the same rules, it is inevitable that different gene-prediction programs would have varying degrees of success. The concurrent application of a combination of methods on a given sequence should circumvent the limitations of individual methods (Fickett, 1996).

It has often been emphasized [see, for example, Fickett (1996)] that the gene sample available in current databases is perhaps atypical, and this can affect the performance of gene-finding algorithms. In this regard, an advantage of the method proposed here is that it does not require a training set, which makes it independent of new sequences being added to databases. The measure exploited in this work is universal, and should prove useful as new and unusual organisms are studied and novel sequences become available for analysis.

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.Subba Rao, C.Chakravarty, D.Searls and R.Guigó 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 Wentian Li for helpful suggestions and participants at the workshop for discussions.

References

Ayres,M.D., Howard,S.C., Kuzio,J., Lopez-Ferber,M. and Possee,R.D. (1994) AcNPV complete sequence. *Virology*, **202**, 586–605.

Bilofsky,H.S. and Burks,C. (1988) The GenBank genetic sequence data bank. *Nucleic Acids Res*., **16**, 1861–1864.

Borodovsky,M. *et al*. (1986) *Mol. Biol*., **20**, 833–840.

Borodovsky,M., Koonin,E.V. and Rudd,K.E. (1994) New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem. Sci*., **19**, 309–313.

Buldyrev,S.V. *et al*. (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E*, **51**, 5084–5094.

Burset,M. and Guigó,R. (1996) Evaluation of Gene Structure prediction programs. *Genomics*, **34**, 353–367. (Sequences for analysis were obtained from the database maintained at the URL http://www.imim.es/GeneIdentification/Evaluation/Index.html.)

Chechetkin,V.R. and Turygin,A.Y. (1995) Size dependence of three-periodicity and long range correlations in DNA sequences. *Phys. Lett. A*, **199**, 75–80.

Dong, S. and Searls, D.B. (1994) Gene structure prediction by linguistic methods. *Genomics*, **23**, 540–551.

Farber,R.B., Lapedes,A.S. and Sirotkin,K.M. (1992) Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol*., **226**, 471–479.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*., **10**, 5303–5318.

Fickett,J.W. (1996) The Gene Identification Problem: An overview for developers. *Comput. Chem*., **20**, 103–118.

Fickett,J.W. and Tung,C.L. (1992) Assessment of protein coding measures. *Nucleic Acids Res*., **20**, 6441–6450.

Fleischmann,R.D. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd’’. *Science*, **269**, 496–512.

Johnston,M. *et al*. (1994) Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science*, **265**, 2077–2082.

Lapedes,A.S., Barnes,C., Burks,C., Farber,R.M. and Sirotkin,K.M. (1990) Application of neural networks and other machine learning algorithms to DNA sequence analysis. In Bell,G. and Marr,T. (eds), *Computers and DNA*. Addison-Wesley, Redwood City, CA.

Li,W. (1992) Generating non-trivial long-range correlations and 1/*f* spectra by duplication and mutation. *Int. J. Bif. Chaos*, **2**, 137.

Li,W., Marr,T.G. and Kaneko,K. (1994) Understanding long-range correlations in DNA sequences. *Physica D*, **75**, 392–416.

Mantegna,R.N., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Peng,C.-K., Simmons,M. and Stanley,H.E. (1994) Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett*., **73**, 3169–3172.

Oliver,S.G. *et al*. (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.

Ossadnik,S.M. *et al*. (1994) Correlation approach to identify coding regions in DNA sequences. *Biophys. J*., **67**, 64–70.

Peng,C.-K. *et al*. (1993) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.

Peng,C.-K. *et al*. (1993) Finite-size effects on long-range correlations: implications for analysing DNA sequences. *Phys. Rev. E*, **47**, 3730–3733.

Sehgal,D., Mittal,V., Ramachandran,S., Dhar,S.K., Bhattacharya,A. and Bhattacharya,S. (1994) Nucleotide sequence organization and analysis of the nuclear ribosome DNA circle of the protozoan parasite *Entamoeba histolytica*. *Mol. Biochem. Parasitol*., **67**, 205–214.

Shulman,M.J., Steiberg,C.M. and Westmoreland,B. (1981) The coding function of nucleotide sequences can be discerned by statistical analysis. *J. Theor. Biol*., **88**, 409–420.

Silverman,B.D. and Linsker,R. (1986) A measure of DNA periodicity. *J. Theor. Biol*., **118**, 295–300.

Synder,E.E. and Storma,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol*., **258**, 1–18.

Tsonis,A.A., Elsner,J.B. and Tsonis,P.A. (1991) Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol*., **151**, 323.

Uberbacher,E.C. and Mural,R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA*, **88**, 11261–11265.

Voss,R.F. (1992) Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. *Phys. Rev. Lett*., **68**, 3805–3808.

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

Xu,Y., Mural,R.J. and Uberbacher,E.C. (1994) *Comput. Applic. Biosci*., **10**, 613–623.

Zhurkin,V.B. (1981) Periodicity in DNA-primary structure is defined by secondary structure of the coded protein. *Nucleic Acids Res*., **9**, 1963–1971.

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