Artificial Intelligence in Medicine (2009) **45**, 117—123

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Detecting conserved coding genomic regions through signal processing of nucleotide substitution patterns

Matteo Ré, Giulio Pavesi *

Department of Biomolecular Science and Biotechnology, University of Milan,
via Celoria 26, 20133 Milan, Italy

Received 15 November 2007; received in revised form 29 July 2008; accepted 30 July 2008

KEYWORDS
Gene annotation;
Gene finding;
Fourier transform;
Comparative genomics

Summary

Objective: In the last few years several complete genome sequences have been made available to the research community. The annotation of their complete inventory of protein coding genes, however, has been so far an elusive goal. Classical ab initio gene prediction methods have been of great support for this task, but show notable weakness in the prediction of genes with unusual structural features. On the other hand, annotation on the basis of similarity to already known genes in other species does not permit the detection of genuinely novel genes and also introduces a potential source of classification error when based on similarity to sequences erroneously annotated as protein coding. Finally, several methods for the functional classification and assessment of evolutionarily conserved regions have been proposed, but, to our knowledge, signal processing techniques have not been applied yet to this problem, despite their proven usefulness at the single genome level.
Results: In this article we introduce the use of signal processing in comparative genomics and we propose a simple test able to evaluate the coding potential of a pairwise genomic sequence alignment according to the pattern and periodicity with which substitutions and gaps appear in the alignment. We assess the feasibility of our approach on an annotated set of human—mouse genomic alignments.
Conclusion: Results show that the application of signal processing techniques to sequence alignments can be a useful tool for the identification of evolutionarily conserved protein-coding regions.

1. Introduction

During evolution, genomic regions under no selective pressure are progressively saturated of mutations,

* Corresponding author. Tel.: +39 02 50314880; fax: +39 02 50315044.
E-mail addresses: matteo.re@unimi.it (M. Ré), giulio.pavesi@unimi.it (G. Pavesi).

0933-3657/$ — see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.07.015

whereas homologous regions under selective pressure retain higher levels of identity. The identification of sequences under evolutionary constraints through the comparison of DNA sequences is a powerful technique for inferring the locations of functional elements in a genome. As whole-genome sequencing efforts extend beyond traditional model organisms to include a wide diversity of species, comparative analyses will be further empowered to reveal insights into genome evolution. The discovery and annotation of functional genomic elements is a necessary step toward a detailed understanding of genome biology, and sequence comparison have been demonstrated to be an integral tool for this task. In recent years an ever increasing amount of findings have clearly hinted that, despite initial assumptions, a large proportion of the sequences conserved between related genomes does not represent protein-coding regions. Other experiments also demonstrated that the classical opinion that long stretches of conserved genomic sequences are predominantly protein-coding regions has to be revised, because of the presence of long conserved non-coding functional elements like modular clusters of well conserved transcription factor binding sites [1]. All in all, the discrimination between conserved coding and non-coding sequences remains an important objective in comparative genomics.

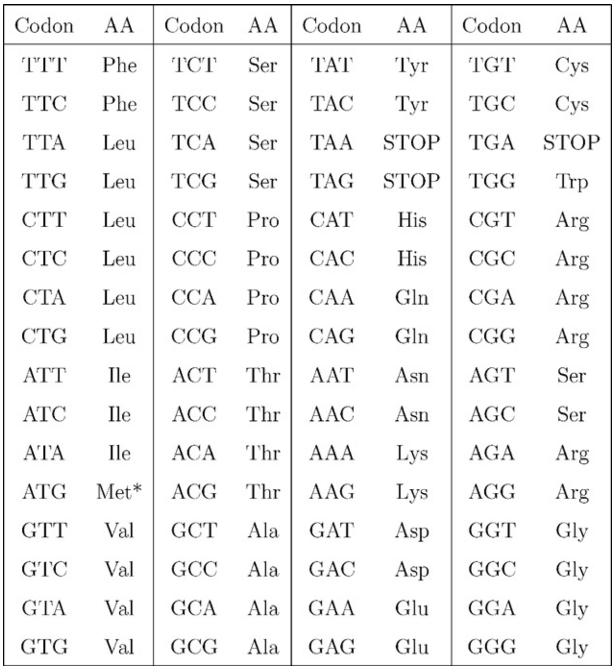
In any comparative analysis, it is of critical importance the accurate choice of the evolutionary distance separating the genomes to be compared. The evolutionary distance separating human and mouse place this pair of organisms at a strategic position for the identification of shared functionally conserved sequences. Around 80 million years (MYs) separate human and mouse from their last common ancestor (160 MYs of independent evolution of their genomes) and the estimated rate of divergence of independently evolving vertebrate genomes, on average 0.1—0.5% per MY, ensures that human—mouse genome comparisons allow sequences whose functional importance is conserved to be identified through sequence alignments, while the evolutionary distance is sufficient to allow the 'masking' of the non-functional ones. On the other hand, at the DNA level even functionally equivalent regions can be difficult to align between more distant species (as it is in the case of human/fish or human/fly comparisons).

Several approaches for protein coding gene prediction based on comparative genomics have been proposed, ranging from TWINSCAN [2], an extension of the ab initio gene predictor GENSCAN [3] that integrates sequence conservation in the probabilistic model (GHMM) of GENSCAN, to CSTminer [4], a tool to discriminate between coding and non-coding conserved sequences on the basis of the presence (or absence) of evolutionary dynamics compatible with a protein coding function.

On the other hand, another class of methods for the detection of protein coding regions are based on the analysis of DNA periodicities, exploiting techniques developed in the field of digital signal processing [5—10]. Despite the fact that these methods have been proven to be quite efficient on a single genome [11], as of today they have not been applied to large scale classification of aligned genomic sequences. Here we present a method based on signal theory that given two aligned conserved genomic sequences classifies them as coding or non-coding according to the pattern and periodicity with which substitutions and gaps appear in the alignment. Indeed, the idea of using substitution patterns in aligned DNA sequences to discriminate protein-coding conserved regions was first introduced in [12], where it was applied to the analysis of a single *Drosophila melanogaster* gene compared its homologs in other *Drosophila* species in order to extend and determine correctly its protein coding sequence. In this work, we extended this basic idea also to deal with frameshifts induced by gaps in the alignments, and tested accuracy and sensitivity of the resulting method on a large set of pariwise human/mouse alignments of coding sequences and intergenic regions, with very encouraging results.

2. Numerical encoding

The periodic pattern in protein coding DNA sequences is a well-known phenomenon. The prominent signal detectable only in protein coding regions, often referred to as "3-periodicity", is a direct consequence of their functional role. In order to produce a new protein a flow of information has to be established from the DNA sequence to the cellular machinery responsible for protein synthesis (the ribosomes). DNA can be seen as a string of symbols belonging from a 4-letter alphabet. In order to encode for 20 amino acids the DNA has to be read in words of length 3 (the codons), and thus there are 64 (4^3) possible codons in DNA, 3 of which are used to encode the end of the translation (protein building) process. The set of rules allowing tRNAs to pair each of the 64 possible codons with the appropriate amino acid are known, in their complex, as the genetic code.



Codon	AA	Codon	AA	Codon	AA	Codon	AA
TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	STOP	TGA	STOP
TTG	Leu	TCG	Ser	TAG	STOP	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met*	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Figure 1 The genetic code: the 64 possible codons, and the corresponding amino acid (AA). ATG serves both as methionine codon and translation start codon; STOP indicates codons marking the end of translation.

Since 20 different amino acids are encoded by 61 codons, the genetic code is redundant, and the same amino acid can be encoded by different codons (see Fig. 1). This is the key point to explain the origin of the 3-period: in families of codons encoding for the same amino acid, each member is used in protein coding DNA regions with different frequencies, leading to a codon usage pattern that is extremely specific for each organism.

The presence of overrepresented codons in coding sequences results into an unequal usage of the four nucleotides in the three positions of the codons and this, in turn, results in a spectral peak at period 3 (because of the codons' size) clearly detectable only in coding sequences and absent in non-coding regions.

The strength of the peak at frequency 1/3 can be easily quantified using Fourier transformation and evaluating the signal over noise ratio in the power spectrum of the DNA sequence under investigation.

At primary level, a DNA sequence *S* [i] of length *N* consists of a series of symbols belonging to an alphabet $\Sigma = \{A, C, G, T\}$. In single sequence signal processing techniques the sequence is mapped to four binary signals, each of which is associated with a specific nucleotide [13]. For example the DNA sequence

$\mathtt{S[i]} = [\mathtt{ATGCGTACGCACTGACGC}]$

can be encoded as follows:

$\mathtt{A[i]} = [\mathtt{100000100010001000}]$

$\mathtt{C[i]} = [\mathtt{000100010101000101}]$

$\mathtt{G[i]} = [\mathtt{001010001000010010}]$

$\mathtt{T[i]} = [\mathtt{010001000000100000}]$

that is, with binary vectors indicating the presence (1) or absence (0) of each nucleotide in each position of the sequence.

Once indicated the discrete Fourier transform (DFT) of the signal associated to each nucleotide (e.g. A) as $\hat{A}(k)$, with $0 \leq k \leq N-1$, the spectral energy associated with sequence S[i] can be defined as follows:

$$|\hat{S}(k)|^2 = |\hat{A}(k)|^2 + |\hat{C}(k)|^2 + |\hat{G}(k)|^2 + |\hat{T}(k)|^2 \qquad (1)$$

Then, for the 3-periodicity property, in protein coding regions the spectral energy obtained by the DFT of the binary signals associated to each nucleotide shows a peak at discrete frequency $N/3$. This peak is not observed in the spectral energy of non-coding DNA regions.

Instead of single sequences, in comparative genomics the objects of investigation are usually aligned sequences. The pattern with which substitutions appear in the alignment can be reasonably expected to provide information regarding the coding potential of conserved sequences. The reason is that the degeneracy of the genetic code tends to make substitutions more tolerated if they occur in the third position of codons, the one where they are less likely to result in a variation of the encoded amino acid, thus maximizing the preservation of the biological function of the encoded polypeptide.

This, in turn, introduces a preferential substitution pattern that can be detected using methods able to quantify the periodicities in signals (such as the DFT). For this reason, the frequency expected to provide maximal discrimination power between coding and non-coding conserved sequences is thus frequency $N/3$, where N is the length (number of columns) of the alignment. For the alignment

```
Aquery[i]       =       [A T G A C T A A G A G A G A T C C G G]
                         | | | | |   | |   | |   | |   | |
Atarget[i]      =       [A T G A C G A A A A G C G A G C C T A]
```

we can build a binary descriptor defining the position of all the substitutions along the aligned sequences:

$$\mathtt{M[i]} = [\mathtt{000001001001001001 1}]$$

To look for aligned regions with mismatches occurring mainly in the third position of a codon, as in [14], we can use the position count function (PCF) to count the number of 1's occurring at each phase $s = \{0, 1, 2\}$ in the binary descriptor M parsed in non-overlapping words of size $w = 3$:

$$C_3^M(s) = \sum_{i=0}^{(N-1)/3} M[3i+s] \quad (2)$$

Using the PCF, as shown in [14] the magnitude of the DFT $\hat{M}[k]$ at discrete frequency $N/3$ can be defined as

$$\left|\hat{M}\left[\frac{N}{3}\right]\right|^2 = \frac{1}{2}[(C_3^M(0) - C_3^M(1))^2 + (C_3^M(1) - C_3^M(2))^2 + (C_3^M(2) - C_3^M(0))^2] \quad (3)$$

Once calculated the signal strength at frequency $N/3$ we need to normalize it with respect to the average spectral noise. The average value $|\hat{M}_{\text{av}}^{(1)}|$ of the squared magnitude $|\hat{M}[k]|^2$ of a binary descriptor, excluding the fundamental frequency component $\hat{M}[0]$, can be calculated as in [14]:

$$|\hat{M}_{\text{av}}^{(1)}|^2 = \frac{1}{(N-1)}\left(N - \sum_{s=0}^{w-1} C_w^M(s)\right)\sum_{s=0}^{w-1} C_w^M(s) \quad (4)$$

where $w = 3$. Finally, the signal (at frequency $N/3$) to noise ratio in the power spectrum representing the spectral coding potential (SCP) of the conserved sequence under investigation can be calculated using the following equation:

$$\text{SCP} = \frac{|\hat{M}[N/3]|^2}{|\hat{M}_{\text{av}}^{(1)}|^2} \quad (5)$$

That is, we measure the coding potential of the alignment as the signal at frequency $N/3$ normalized with respect to the signal present at every frequency of the spectrum excluding the dc component ($\hat{M}[0]$).

A potential problem affecting the SCP is represented by frame shifts introduced by gaps in aligned sequences. A phase shift in a member of a pairwise sequence alignment leads to an enrichment of substitutions in a wrong phase and this might result in an artifactual increase of the three periodicity in the substitution pattern. The problem can be solved considering the differences in the pattern of gaps observable in protein-coding and non-coding aligned sequences. In alignments of coding sequences gaps often occur in multiples of three (corresponding with amino acid indels in the translated protein) and, even if the presence of a single gap disrupts the correct phase alignment, the correct phase is recovered by the presence of other gaps nearby, recovering the correct phase. Starting from these observations, metrics able to calculate the amount of columns in the alignment representing different pairs of aligned phases are highly informative on the functional nature of the conserved regions, since a pair of coding aligned sequences is expected to show a very strong phase conservation as opposite of non-coding ones.

In our experiments we used a phase conservation metric in order to protect the calculated coding potential score from 3-periodic substitution patterns randomly introduced by frame shifts introduced by gaps in the aligned sequences. As frame-recovery test we adopted the technique proposed by Noguchi and colleagues [15]. Possible frame shifts can be classified with respect to the remainders of gap lengths divided by three. There are thus three possible cases. A three-base shift does not introduce misalignments of reading frames, regardless of the location of the gap in either of the sequences aligned. A single base shift caused by a gap in one sequence corresponds to a two base shift in the other, or vice versa (see Fig. 2 A—C). However, the examples illustrated in Fig. 2 are based on prior knowledge of the correct reading frame, which is not available during alignment of unannotated sequences. Hence, in [15] the authors defined an index representing the 'phase lag' between two aligned sequences with three base periods.

According to [15] we thus calculate the phase lag status associated to each nucleotide in both the sequences aligned. Let a (phase — 0, 1 or 2) be the phase in the first sequence at column i in the alignment and b (phase — 0, 1 or 2) the corresponding phase label in the second sequence at the same position. We counted the occurrences of each phase alignment combination (PAC) along the alignment columns and we then calculated its frequency among all the gap-free columns in the alignment.

We then define the frame recovery test (FRT) as the frequency with which the most frequent phase pair is found in the gapless positions of the alignment:

$$\text{FRT} = \arg\max_{(a,b)} \frac{\text{PAC}(a,b)}{\text{gf}_\text{c}} \quad (6)$$

where PAC(a, b) is the phase alignment combination of each possible pair of phases (a, b), and gf_c denotes the number of gap-free columns in the alignment. The use of FRT as part of a coding potential metric is justified by the observation that in coding exons of conserved genes, gaps in the

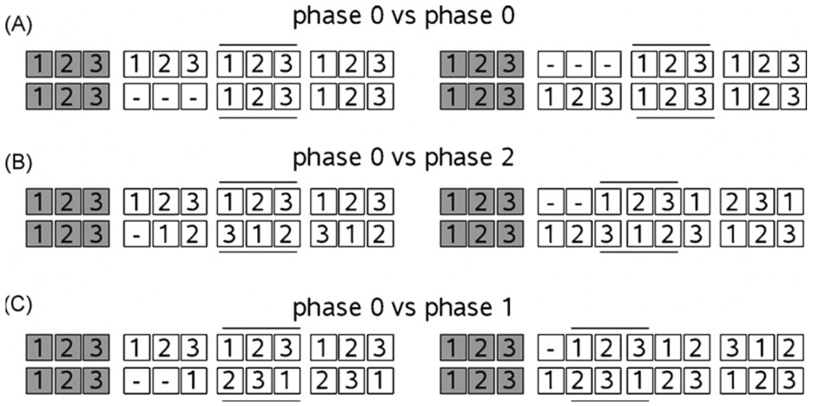


Figure 2 Gaps in aligned sequences and the consequent variation of phase of aligned pair of nucleotides produces three main classes of frameshifts, according to the number of contiguous gaps and their position in the top sequence or in the bottom sequence.

alignment do not shift the reading frame or recover the frames even if they are shifted and, more important, this index can be applied in absence of prior knowledge about the real reading frame of the aligned sequences.

The final coding potential metric we introduce is defined by the multiplication of the SCP (as described by Eq. 5) and the FRT:

$$\mathrm{CP} = \mathrm{SCP} \cdot \mathrm{FRT} \tag{7}$$

3. Experiments

In order to assess the ability of our method to discriminate between coding and non-coding conserved sequences we built two evaluation datasets. The coding dataset contained 3061 alignments obtained comparing 1580 pairs of human and mouse orthologous coding sequences retrieved from Biomart. The simplest way to obtain a non-coding sequence set was to align whole genomic intergenic regions, and then to remove from the alignment sequences overlapping genomic regions annotated as coding. In particular, we employed the human genomic sequences annotated in the ENCODE project, and their alignment with the corresponding homologs in mouse. The high quality of the existing annotations for these regions allowed us to safely discriminate between alignments containing protein coding regions, even if we could not exclude a priori the presence of unknown protein coding genes or pseudogenes. The comparison of ENCODE regions and their homologs in mouse produced 4123 alignments. The comparison of genomic coordinates of the alignments with the content of the ENTREZ and VEGA gene databases led to the removal 1896 alignments, 1771 of which overlapped annotated coding sequences and 125 overlapped to annotated pseudogenes. Thus, the resulting non-coding set was composed by 2227 aligned sequence pairs. The alignments (coding and non-coding) ranged in length from 42 to 1032 base pairs, with no significant difference in the length distribution of coding and non-coding sets.

We further improved the quality of our alignments set removing all the alignments shorter than 60 nucleotides and with less than 5% of divergence at nucleotide identity level. This is expected to provide optimal conditions for our tests, since we are trying to quantify the amount of differential variation observable in different codon positions and thus we need to grant the existence of both a minimal amount of variation and a relatively high number of codons.

The final positive set (coding set) comprised 2929 alignments while the negative set (non-coding set) was composed by 1779 pairwise alignments.

Each alignment was analyzed in order to measure the signal over noise ratio introduced at frequency $N/3$ by substitutions occurring in the third position of codons corrected by a term accounting for the frame shift recovery due by the non-random position and density of gaps in alignments protein coding conserved sequences. The CP index calculated according to Eq. (7) ranged from 0.00 to 612.09 in the 2929 alignments composed only by coding sequences, and from 0 to 35.12 in the 1779 sequences of the non-coding set. The cumulative distribution of CP values in coding and non-coding sets is shown in Fig. 3. As we can see, CP is highly discriminative between the alignments contained in our benchmark sets.

Table 1 shows the number of coding sequences correctly classified (true positives), and uncorrectly classified as non-coding (false negatives), and vice versa for non-coding (false positives and true negatives), with the corresponding sensitivity and

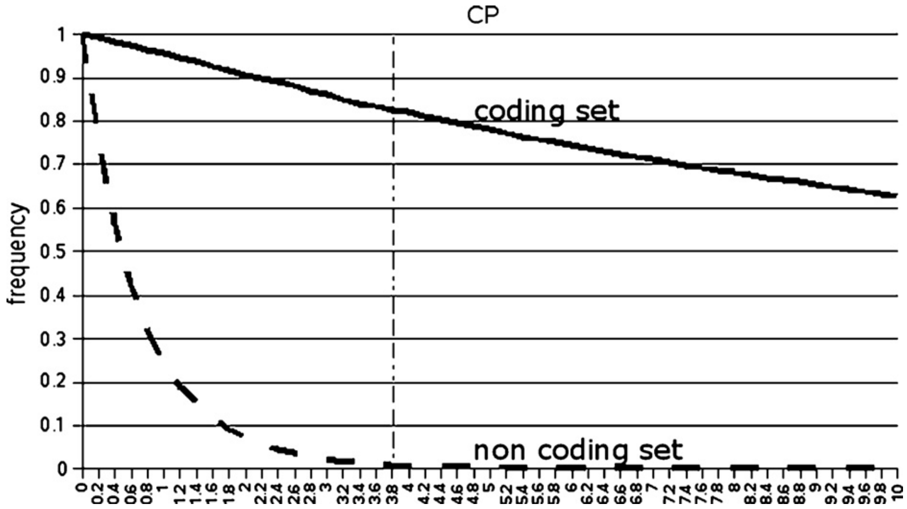


Figure 3 Fraction of coding (continuous line) and non-coding alignments (dotted line) with CP score greater than the *x*-axis value. A CP cutoff set to 3.8 allows for the correct prediction of 83% of the positive (protein coding) set with a false positive rate of 1.0%.

Table 1 At different CP score threshold values, fraction of alignments correctly and uncorrectly classified as coding (true positives—TP, false positives—FP), and correctly and uncorrectly classified as non-coding (true negatives—TN, false negatives—FN), and the corresponding sensitivity (TP/(TP + FN)) and specificity (TN/(TN + FP)) values

CP	TP	FP	TN	FN	Sens	Spec
0.0	1	1	0	0	1.00	0.00
2.0	0.91	0.07	0.93	0.09	0.91	0.93
4.0	0.82	0.01	0.99	0.18	0.82	0.99
6.0	0.74	0.00	1.00	0.26	0.74	1.00

specificity values at different CP threshold values. For example, the CP score at threshold 4.0 is able to classify correctly 2405 coding alignments (82.10% of coding set) yielding only 12 false positives (less than 1.0% of the non-coding set).

Further examination of false positive alignments revealed that seven of the false positives obtained by the method matched transcribed regions (that is, annotations like RNAs, cDNAs, ESTs), while four more overlapped proteic features (like conserved domains) indicating the possible presence of as yet unannotated genes in nearly all of our "false positive" predictions.

4. Conclusions

The periodicity of three detectable at nucleotide level in coding regions has been observed by many authors, even if spectral techniques derived by this observation, to our knowledge, have never been applied to a comparative analysis. In this paper we presented the application of spectral techniques to aligned sequences and we demonstrated that the signal over noise ratio at discrete frequency $N/3$ (where N is the length of the alignment) obtained transforming a binary indicator encoding the positions of substitutions can be effectively used for discrimination between protein coding and non-coding aligned DNA sequences.

This observation is a direct effect of the characteristic selective pressure to which only functional and protein coding conserved regions are subject during evolution. The important problem of periodicities disruption due by the presence of frame shifts in pairwise alignments has been addressed using a relatively simple frame recovery test. The method we presented can be further improved using more complex signal processing approaches and introducing other correction factors.

A straightforward application of our method is the annotation of newly sequenced genomes, because, once defined an appropriate cutoff value, a classification can be obtained in total absence of any previous knowledge regarding the genomic regions under investigation. Because the origins of the signal we investigated in this work is the selective pressure acting on protein coding regions and because this is due to the presence of a near universal genetic code allowing the use of information encoded in DNA for protein synthesis, we expect the method to be valid for investigations in pairs of species other than human and mouse, as well as the analysis of multiple sequence alignments, or the characterization of RNA sequences as coding (mRNAs) or non-coding. Indeed, our method is currently being integrated in a gene prediction tool which will be tested on genome-wide alignments of different species, and made available to the research community.

References

[1] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS. Ultraconserved elements in the human genome. Science 2004;364(5676):1321—5.

[2] Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics 2001;17(Suppl. 1):S140—148.

[3] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268(1):78—94.

[4] Mignone F, Grillo G, Liuni S, Pesole G. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucl Acids Res 2003;31(15):4639—45.

[5] Anastassiou D. Frequency-domain analysis of biomolecular sequences. Bioinformatics 2000;16(12):1073—81.

[6] Issac B, Singh H, Kaur H, Raghava GP. Locating probable genes using Fourier transform approach. Bioinformatics 2002;18(1):196—7.

[7] Kauer G, Blocker H. Applying signal theory to the analysis of biomolecules. Bioinformatics 2003;19(16):2016—21.

[8] Kotlar D, Lavner Y. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res 2003;13:1930—7.

[9] Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 1997;13(3):263—70.

[10] Yan M, Lin ZS, Zhang CT. A new Fourier transform approach for protein coding measure based on the format of the z curve. Bioinformatics 1998;14(8):685—90.

[11] Vaidyanathan P, Yoon BJ. The role of signal-processing concepts in genomics and proteomics. J Franklin Inst 2004;341: 111—35.

[12] Tatarenkov A, Saez AG, Ayala FJ. A compact gene cluster in Drosophila: the unrelated Cs gene is compressed between duplicated amd and Ddc. Gene 1999;231(1—2):111—20.

[13] Voss R. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 1992;68:3805—8.

[14] Datta S, Asif A. A fast DFT based gene prediction algorithm for identification of protein coding regions. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing ICASSP 2005. Philadelphia: IEEE Press; 2005. p. 113—6.

[15] Noguchi H, Yada T, Sakaki Y. A novel index which precisely derives protein coding regions from cross-species genome alignments. Genome Inform 2002;13:183—91.