J Mol Evol (2004) 59:309–316
DOI: 10.1007/s00239-004-2626-7

JOURNAL OF MOLECULAR EVOLUTION

Measuring the Coding Potential of Genomic Sequences Through a Combination of Triplet Occurrence Patterns and RNY Preference

Christoforos Nikolaou, Yannis Almirantis

Institute of Biology, National Research Center for Physical Sciences "Demokritos," 15310, Athens, Greece

Received: 15 July 2003 / Accepted: 10 March 2004 [Reviewing Editor: Dr. Massimo Di Giulio]

Abstract. The distribution of *n*-tuplet frequencies is shown to strongly correlate with functionality when examining a genomic sequence in a reading-frame specific manner. The approach described herein applies a coarse-graining procedure, which is able to reveal aspects of triplet usage that are related to protein coding, while at the same time remaining species independent, based on a simple summation of suitable triplet occurrences measures. These quantities are ratios of simple frequencies to suitable mononucleotide-frequency products promoting the incidence of the RNY motif, preferred in the most widely used codons. A significant distinction of coding and noncoding sequences is achieved.

Key words: Triplet occurrence — Coding potential — RNY preference

Introduction

Data from completed and ongoing genome projects are rapidly accumulating. Under this prospect, the need for efficient tools for analyzing specific properties of the newly found sequences is becoming imperative. Methods formulated for this task fall mainly in two categories (Fickett 1996). "Signal" methods make use of nucleotide strings that are highly correlated with specific functions of the genomic material, such as splice junctions, consensus sequence elements of promoter regions, etc. "Content" methods, on the other hand, are based on statistical features of the coding genetic text as contrasted with the noncoding from oligonucleotide frequencies up to higher-order statistical properties (Fickett and Tung 1992; Rogic et al. 2001).

A key aspect for the study of the genomic text is the length scale, under which the sequences are examined. A DNA primary structure can be studied under several length scales, each providing a "filter," through which specific statistical attributes of the sequence, are revealed. In this work we focus on the short scale, the one lying below 10 nucleotides, which is immediately affected by the grammar and syntax of the genetic code.

When examining this particular length scale, studies of the patterns of oligonucleotide (*n*-tuplet) occurrences are of critical importance. Deviations from randomness in the *n*-tuplet occurrence have been extensively studied using several methodological approaches (Burge et al. 1992; Gutierrez et al. 1993; Karlin and Ladunga 1994). The observed patterns have been directly visualized, through approaches like the Chaos Game Representation (Jeffrey 1990) and "genomic portraits" (Hao 2000a, b). In general, the observed patterns have been found to be species or taxonomic group specific, allowing the derivation of evolutionary trees similar to those obtained using other approaches (Karlin et al. 1994; Karlin and Mrazek 1997). On the contrary, in most cases, the nucleotide *n*-tuplet occurrences are not clearly cor-

Correspondence to: Yannis Almirantis; *email:* yalmir@bio.demokritos.gr

related with the functional role of the examined sequences (Burge et al. 1992; Karlin and Burge 1995; Nussinov 1981). Whenever such statistical differences are found between coding and noncoding segments, they are too weak or superimposed with coexisting species-specific patterns and therefore cannot be used in tools for the detection of new protein-coding regions.

Three nucleotide long words are of particular interest for obvious reasons, having to do with the nature of the genetic code. The use of triplets is essential in the case of coding sequences. Optimization of the frequencies of trinucleotides in order to achieve high expression fidelity and speed is a common strategy, used by a wide range of organisms, known as codon usage (Bulmer 1991; Sharp and Li 1986). In addition, codon usage patterns are widely used in phylogenetic studies, as well as in attempts to estimate the expression rate of a given gene. Therefore we see that the nonrandom usage of codons is considered as a strong indication of the coding potential of a sequence, and in most cases it is also correlated with its expression rate. A number of studies have pointed out the need to account for background nucleotide composition when studying codon usage (Akashi and Eyre-Walker 1998; Akashi et al. 1998; Kliman and Eyre-Walker 1998; Marais et al. 2001; Novembre 2002; Urrutia and Hurst 2001). This is exceptionally imperative in the cases of higher eukaryotic genomes, where nucleotide composition is subject to large intragenomic fluctuations (Bernardi 1989; Bernardi 1993).

The approach described in the present work is based on a reading-frame specific counting of frequencies of triplet occurrences, which then are normalized over a suitable mononucleotide-frequency product promoting the incidence of the RNY motif. This division ascribes a statistical weight to the value of each observed frequency of occurrence. Then the final quantity is obtained by a simple summation of measures of *n*-tuplet frequencies, that being a coarse-graining procedure. A suppression of species-specific features in the triplet distribution is achieved, thus revealing characteristics of the sequence having to do with its coding role. It is therefore expected to be able to distinguish systematically between coding and noncoding sequences.

Sequences and Data Handling

Collections of sequences of known origin, functionality, and mean lengths were downloaded from EMBL database using the SRS7 retrieval system in the following way.

Large species-specific collections, each including all sequences of a given origin and functionality and within a given length range, were initially formed. From each of those raw collections, 500 entries randomly chosen were retrieved, thus resulting in collections with minimized redundancy, which were the final objects in our analysis.

Collections of mixed origin were also formed using the EMBL database. In this case the interest focused on sequence representation from more general categories, namely, higher eukaryotes, viruses, and organelles. The eukaryotic collections consisting of coding sequences (CDS) and intronic sequences, with mean length ~4000 nucleotides (~4 knt), were completely cleaned for redundancy, thus constituting the most reliable reference set. Nevertheless, sequence collections of lower mean lengths, not checked for redundancy, behaved in all cases in a manner similar to that of the nonredundant set, a strong indication of the very slight impact that sequence redundancy has on the obtained results.

Prokaryotic and yeast coding and noncoding sequence collections were obtained from 30 complete representative eubacterial genomes and the 16 chromosomes of *S. cerevisiae*, respectively, as retrieved from GenBank.

Trying to quantify the success of several algorithms, based on quantities measuring "coding potential" in separating sets of (known functionality) coding and noncoding sequences, we proceed as follows.

Having the two sets of "test" sequences represented by two distribution curves of the given quantity (Q), we first determine numerically an optimal threshold value Q_{thr}, which divides the Q-value space into two separate regions, one hosting mainly coding and the other noncoding sequences. Accordingly, the "test" sequence sets are divided into four subpopulations: True and false coding and noncoding sequences (TC, FC, TN, and FN, respectively). Then we define as "classification rate" the ratio (TC + TN)/(TC + TN + FC + FN) expressed as a percentage. Notice that the collections to be compared were always chosen to contain sequences of equal mean length and originating from the same species or species group.

The Codon Occurrence Measure

Method

The algorithm used for the computation of Codon Occurrence Measure (*COM*) is described below.

(A). The whole sequence is read in a reading frame (RF)-specific context, calculating the 64 trinucleotide measures of occurrence in various ways, which are described later. Three different values of what we call the codon occur-

rence measure (COM) are obtained, by summing up the calculated measures of occurrence for each of the three reading frames individually:

$$COM_{RF} = \sum_{64} R_{ijk}(i,j,k = A,G,C,T) \quad (RF = 1,2,3)$$

In the above formula R_{ijk} designates the used measure of occurrence of triplets (see below for variations of its definition) for each RF.

(B). The maximum of these three values is taken to be the COM.

$$COM = max\ (COM_1,\ COM_2,\ COM_3)$$

The method of calculation of R_{ijk} is of great importance. The summation of all simple frequencies of occurrence, by definition, gives a total equal to unity. Only considering ''normalized'' quantities could provide information relevant to compositional preferences or avoidances, related to the coding character of the examined sequence.

- One may use odds-ratio frequencies over the corresponding mononucleotide frequencies used extensively in the literature (Blaisdell 1986; Brendel et al. 1986; Stuckle et al. 1990, 1992). These are calculated through division of the simple triplet frequencies of occurrence F_{ijk} over the product of the frequencies of occurrence of the constituting mononucleotides F_i, F_j, F_k. Deviations of the odds ratios from unity measure the over- / under-representation of trinucleotides from expected values, estimated using a zeroth-order Markov process:

$$R_{ijk} = F_{ijk}/F_iF_jF_k$$

- A further refinement could be the use of position-specific mononucleotide frequencies of occurrence according to the formula:

$$R_{ijk} = F_{ijk}/F_{i(1)}F_{j(2)}F_{k(3)}$$

Here the subscripts (1), (2), and (3) designate the position of the mononucleotide since the mononucleotide frequencies are also computed in a frame-specific manner, $F_{i(1)}$ meaning the frequency of nucleotide i in the first codon position for the reading frame examined and accordingly for the second and third codon positions.

- Furthermore, a modification of the odds ratio, previously introduced in a study of the asymmetry of DNA sequences (Nikolaou and Almirantis 2003), may be used. This modification is based on the observation (Crick et al. 1976) that highly used codons in all species generally tend to be of the form RNY (R, purine: Y, pyrimidine: N, any base). RNY codons are widely used and comprise the fraction of the most preferred codons in all known organisms. This specific preference has been attributed to various reasons, either the existence of an ancient genetic code (Eigen and Schuster 1977) or selection due to evolutionary advantages (Hanai and Wada 1989; Wong and Cedergren 1986), and has been used in methods of determination of the correct reading frame (Shepherd 1981, 1990). The RNY pattern introduces an additional asymmetry inside the codons. In this context, the codon structure factor (CSF), which incorporates the observed mononucleotide frequencies specific of position and in reverse order of the one implied by the examined triplet, is introduced. CSF is incorporated in the calculated triplet frequencies of occurrence as follows:

$$R_{ijk} = R_{CSF} = F_{ijk}/CSF_{ijk} = F_{ijk}/F_{i(3)}F_{j(2)}F_{k(1)}$$

In this way, in a coding sequence, any triplet following the ''RNY rule'' will be subsidized since the division will be over an inferior denominator, while, on the other hand, triplets deviating from the above ''rule'' will be attributed a lower R_{ijk} value. In this way, sequences exhibiting high COM values (when CSF is incorporated) are very likely to be coding and probably represent genes with high expression rates.

Table 1 shows the classification rates obtained using various modifications of COM. Odds ratios using ''position-independent'' single-nucleotide frequencies yield the poorest results (practically no distinction). This seems reasonable, considering that a significant percentage of the informative ''load'' carried by the sequence is lost when dividing over single-nucleotide frequencies. That is because the mononucleotide products used as denominators carry meaningful information concerning amino acid composition and codon usage skews. In this way, division over these quantities reduces the amount of information with which the sequence is endowed. Odds ratios computed individually over the three codon positions slightly improve the obtained classification rates. This could be explained taking into account the use of position-dependent mononucleotide frequencies of occurrence, which occasionally reach lower values than position-independent ones, due to codon biases in coding sequences. This tends to drive the formed fractions to higher values, thus contributing to a systematic increase in COM for coding sequences if compared to noncoding ones. Notice that this explanation is based on the fact that unevenness of mononucleotide frequencies of occurrence drives ratios having them in the denominator to systematically increase.

Simple division over the codon structure factor gives, overall, the best results. Alternative variations of the form of CSF-triplet occurrence measures, the

Table 1. Assessment of several COM method variations through classification rates for mixed-origin sequence collections of various mean lengths

	Mode of triplet-occurrence estimation				
	Odd ratios				
Set of comparison	Position independent	Position specific	R_{CSF}	$\lvert R_{CSF} - 1\rvert$	$(R_{CSF} - 1)^2$
Eukaryotic					
CDS/introns ~500 nts	54.0%	60.7%	86.9%	87.2%	**87.9%**
CDS/introns ~1000 nts	50.1%	59.0%	**90.7%**	91.0%	90.7%
CDS/introns ~2000 nts	50.7%	59.8%	**98.0%**	95.3%	94.2%
CDS/introns ~4000 nts	52.4%	61.3%	**96.5%**	94.0%	93.8%
Prokaryotic					
Coding/noncoding ~1000 nts	50.1%	59.5%	**84.7%**	76.9%	78.2%
S. Cerevisiae					
Coding/noncoding ~1000 nts	51.2%	60.3%	94.1%	93.8%	**94.9%**

Note. Maximal classification rates shown in bold support the simplest form incorpating CSF.

results of which are shown in Table 1, were also examined. The quantities $|R_{CSF} - 1|$ and $(R_{CSF} - 1)^2$, where R_{CSF} designates measures of triplet occurrences as defined above, were used. The results overall remain optimal for the case of the simple summation of R_{CSF}. Therefore, this simplest form (Table 1, column 4) was used throughout the subsequent analysis as the one with optimal behavior.

Length and Species Dependence of COM

The methods' dependence on sequence length and species was of primary interest. To test this, two sets of collections of 500 coding sequences each were formed. The first set comprised four collections of eukaryotic CDSs, originating from different eukaryotes, not including fungi and protists and having mean sequence lengths of 500, 1000, 2000, and 4000 nucleotides. This set was used to test the length dependence of the method. The collections' COM-value distributions are shown in Fig. 1. One can easily observe that the distribution curves have very similar mean values and standard deviations (numerical data not shown), being almost completely overlapping. This comes as an indication that the method is affected by the sequence length to only a very small extent, more visible in the shortest sequence collections, having to do with the finite size effect and as a result of poor statistical representation of triplet occurrences of lengths of the order of 500 nt. Collections of intronic sequence in the same length ranges behave similarly (data non shown).

The second set consisted of five discrete CDS collections with the same mean length, ~4 knt, originating from five particular eukaryotic species (*Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, and Saccharomyces cerevisiae*). The COM-value distributions of these collections are depicted in Fig. 2. We see a significant overlapping of the five distribution curves, a fact indicative of the species independence of the *COM* method. This may be understood on the basis of the structure of this method, which takes place through a coarse-graining simple summation of all triplet-occurrence estimates, thus canceling out expressed species-specific patterns and revealing, at the same time, the ones that correlate with the sequence's functionality. In this way, the CDS distributions of species with considerable evolutionary distances are centered around similar mean values and exhibit standard deviations in a close vicinity (data not shown). This leads us to the conclusion that the method is able to capture statistical properties related to the protein coding procedure that are common in a very wide range of species, if not universal.

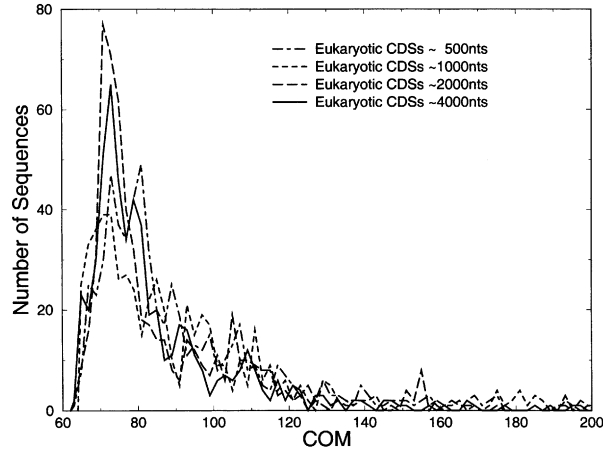

Fig. 1. *COM*-value distributions of eukaryotic coding sequence collections of various mean lengths.

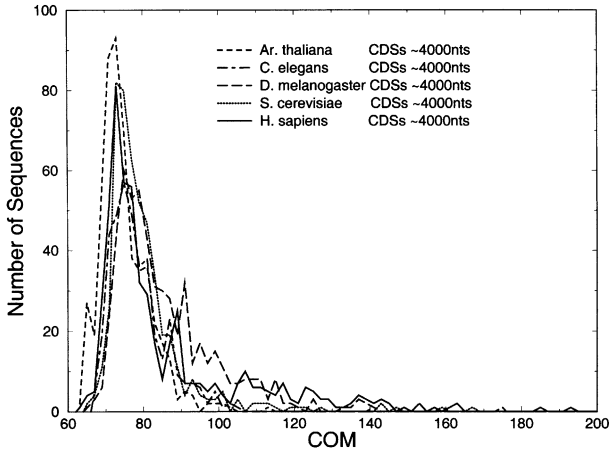

Fig. 2. *COM*-value distributions of coding sequence collections, originating from five different eukaryotic species.

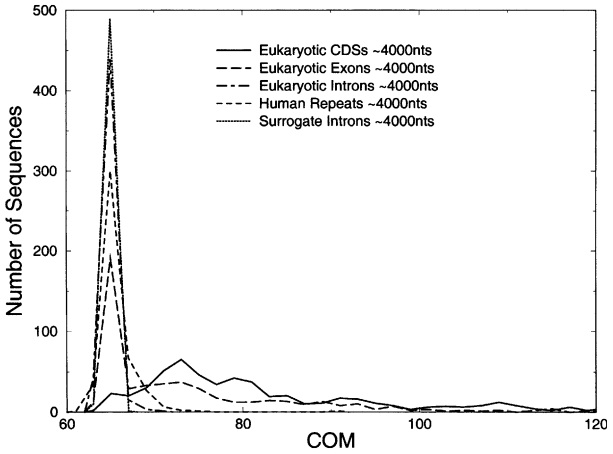

Fig. 3. *COM*-value distributions of eukaryotic coding, intronic, exonic, and repeated sequence collections with a mean length P ~4000 nt. A surrogate sequence collection of the same mean length is also drawn for direct comparison to the behavior expected under randomness. Notice that the maxima for all distributions except coding sequences are located at 64, which is the expected value for random sequences.

Higher Eukaryotic Genomes

We went on to examine the behavior of specific categories of sequences, regarding triplet usage patterns as reflected in *COM* values. Principal components of the genomes of higher eukaryotes are analyzed in Fig. 3. One can observe the clear distinction between coding and noncoding sequences, comparing collections of CDSs versus introns (both of length ~4 knt in this case). One may also notice that the CDS distribution curve is very dispersed and located in high *COM* values. On the other hand, the sharp intronic curve is centered in lower values around the value expected under random distribution (equal to $4^3 = 64$) and clearly overlaps the corresponding surrogate collection distribution. High mean values are representative clues of coding potential since *COM* is positively correlated with nonrandom usage of codons, as would be expected for sequences under coding and translational constraints. Moreover, high values of standard deviation are indicative of the wealth of codon usage patterns among protein coding sequences, representing the great range of different statistical attributes between specific protein families. Noncoding sequence curves lack both high mean and dispersions, as expected for sequences where nucleotides are juxtaposed under random distribution at the very short scale. The overlapping area between the two curves is quite small (~3.5% of the total sum as shown in Table 1, column 4), a positive indication for the *COM* discriminating power.

In Fig. 3, a *COM* distribution curve corresponding to exons with a length around 4 knt is also included. The particular curve clearly consists of two discrete parts. One sharp peak falls in the region of introns of the same length and one long tail, spanning the region typical for the CDSs. This finding is in accord with previous studies on exonic sequences applying different methodological approaches (Nikolaou and Almirantis 2002, 2003). We have checked one by one the total (~41% of the collection) of the sequences contributing to the sharp peak region for the COM curve in Fig. 3. We have found that 87% are terminal exons translated in less than 20% of their length. On the other hand, in the dispersed rightmost part of the curve sharing the general shape of the CDS distribution, such nontranslated exons are present in only 5%. The above observation reveals the *COM* efficiency in reflecting genomic properties having to do with the coding potential of a sequence.

The distribution of a collection of human repeat sequences (with a mean sequence length of ~4 knt) has the sharp shape characteristic of the noncoding sequence but is slightly shifted toward higher *COM* values. This can be justified in terms of the used triplets occurrence. On the one hand, repeated sequences have an implicit overrepresentation of "words" deviating from a random behavior. On the other hand, their repeating primary structure imposes essential constraints that attribute occurrence patterns fundamentally different from the ones expected for coding sequences, as long as the repeated part is not of a length that is a multiple of three. Moreover, even in repetition with a preponderant 3-nt period, their repetitive structure is very likely to be interrupted by small insertion sequences, canceling the effects of the repetition unit in a single reading frame and thus diminishing the *COM* value.

Promoter and rRNA coding sequence collections have also been tested. As expected, due to the lack of protein-coding information in these categories of genomic sequences, their *COM* values fall in the range of noncoding sequences (data not shown).

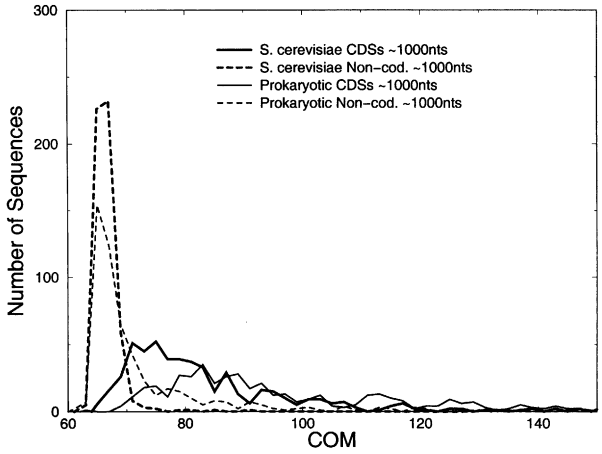

Fig. 4. *COM*-value distributions of coding sequences and noncoding segments originating from prokaryotes (30, eubacterial species) and *S. cerevisiae* with a mean length of ~1000 nt.

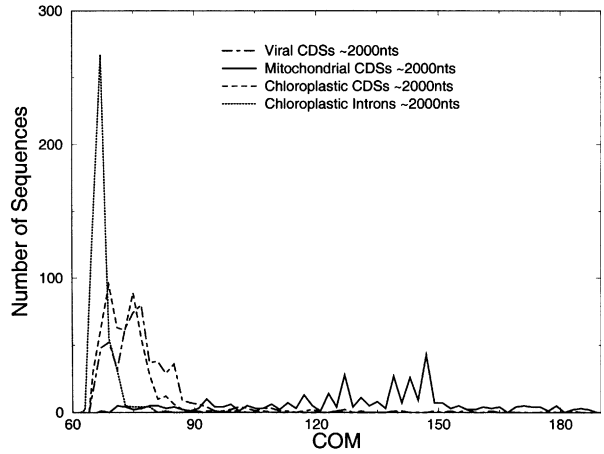

Fig. 5. *COM*-value distributions of coding sequences, originating from viral, chloroplast, and mitochondrial genomes with a mean length of ~2000 nt. The distribution curve of the chloroplast intronic sequence collection of the same mean length is drawn for a direct comparison of coding and noncoding sequences.

Lower-Complexity Genomes

Prokaryotic as well as fungal genomes differ significantly from higher eukaryotic ones in their coding percentage and extent of regulatory and repeating elements, among several other aspects of genome organization. Nevertheless, specific triplet usage patterns have also been observed in simple organisms like prokaryotes (Makhoul and Trifonov 2002).

Collections of prokaryotic and yeast coding and noncoding sequences with a mean length of ~1000 nt were analyzed and the results are represented graphically in Fig. 4. The discrimination between coding and noncoding sequence collection distributions, initially observed for higher eukaryotes, is again present in simpler genomes. The overlapping curve areas are quite narrow, reaching 6% in the case of yeast and 15% for prokaryotes.

A major difference between the higher- and the lower-complexity genomes in terms of *COM*-assisted discrimination is the following. In the prokaryotes, it is the noncoding curve skew that is mainly responsible for the overlap percentage, whereas in higher eukaryotic genomes this situation is inverted. This can be justified taking into account the small percentage of noncoding space in prokaryotic genomes of lower complexity, not exceeding 10% in most prokaryotes. This means that noncoding spacers are almost always in the close neighborhood of coding regions, thus partially retaining some of their specific features, probably due to different positions of the coding/noncoding borders in the evolutionary past.

Parasitic and Symbiotic Genomes

Genomes that maintain symbiotic or parasitic relationships with eukaryotic organisms such as viral, mitochondria, and chloroplastic ones were analyzed next. In Fig. 5 we have drawn the distribution curves of three ~2000-nt-long CDS collections, taken from viral, mitochondrial, and chloroplastic genomes, presented alongside equal-length chloroplast introns. As expected, coding sequences from such genomes share characteristics similar to those of their hosts, having dispersed distribution curves located in the high-*COM* values region, contrasting with intronic sequences originating from chloroplast genomes (the only organelle group able to provide a large number of sequences of this functionality), which are located at lower values. Notice that the organelle sequence collections depicted here present a high degree of redundancy, due to the restricted number of proteins encoded by their genomes.

Distinguishing Coding and Noncoding Sequences Through COM

Finally, we applied the *COM* measure on collections of sequences with different functionality, in order to estimate quantitatively the efficiency of the method in discriminating between coding and noncoding sequences. The results, presented in Table 2, show very high classification rates (calculated as described earlier) for increasing sequence length, while remaining relatively high for eukaryotic sequences down to 500 nt long. Classification rates are generally above 90% for all eukaryotic collections tested. This leads us to the conclusion that the proposed method, combined with already existing algorithms, may provide a useful tool for assessing the coding potential of a sequence.

Discussion

n-Tuplet usage has been strongly correlated with the sequence's origin. Nevertheless, compositional con-

Table 2. Coding/noncoding sequence classification rates obtained from the application of the simplest COM method for sequences of various origins and mean lengths (R_{CSF} COM Variation)

Set of comparison	Classification rate
Eukaryotic	
CDS/introns ~500 nts	86.9%
CDS/introns ~1000 nts	90.7%
CDS/introns ~2000 nts	98.0%
CDS/introns ~4000 nts	96.5%
A. thaliana	
CDS/introns ~1000 nts	88.4%
D. melanogaster	
CDS/introns ~1000 nts	94.3%
H.sapiens	
CDS/introns ~1000 nts	91.3%
CDS/introns ~2000 nts	98.1%
CDS/introns ~4000 nts	99.6%
S. cerevisiae	
Coding/noncoding ~1000 nts	94.1%
Prokaryotic	
Coding/noncoding ~1000 nts	84.7%

straints existent in all known species are observed in both coding and noncoding sequences. These constraints differ between these two classes of functionality and are therefore expressed through different patterns of *n*-tuplet usage distributions. The approach described above is able to capture such differences that remain species independent and consecutively use them to distinct coding from noncoding sequences. Patterns of *n*-tuplet usage are mainly affected by codon and amino acid bias in coding sequences and exhibit great diversity reflecting the wealth of protein structures and functions exhibited in any genome. This particular property of the protein-coding sequences, as reflected by high *COM* values, is able to distinguish between coding and noncoding sequences. One sees that a simple summation of triplet-occurrence measures (here of the R_{ijk} values incorporating the codon structure factor) is able to reveal properties of the sequence that are directly related to its functional role.

The *COM* algorithm incorporates aspects addressed by a variety of other coding model-independent statistics in a simple quantity. For a comprehensive presentation of several such algorithms see Guigó (1999). *COM* is, by construction, directed at measuring triplet occurrences. In this way it is correlated with both codon and amino acid usage while, at the same time, able to capture the underlying 3-nt periodicity (Guttiérez et al. 1994; Tiwari et al. 1997) and the RNY codon pattern (Shepherd 1981). Furthermore, aspects of mutual information on the sequences such as the codon pattern ones used by Herzel and Grosse (1995) are taken implicitly under consideration, as by construction we calculate ratios of triplet occurrences over position-specific mononucleotide frequencies. In this way, apart from expecting an increased efficiency, *COM* is characterized by the incorporation of a variety of attributes used for coding statistics, through a relatively simple calculation.

The *COM* method may potentially serve as an additional estimator for ascribing the functional role of a given sequence. High *COM* values would be in agreement with the coding character of a sequence (predicted by standard gene-finding tools), while low ones would suggest the reassessment of its supposed functionality. *COM* would be suitable for organisms with few known genes, as it does not require extensive training with known sequences. For the same reasons, it fits well to organisms with a high inhomogeneity of genomic constitution. Moreover, *COM*-based gene finding techniques could detect genes transferred "horizontally" in a genome. Usually gene-finders trained with sets of genes of the host genome fail to recognize these genes (Lukashin and Borodovsky 1998; Kraemer et al. 2001). Preliminary results on the combination of *COM* with other "coding potential—specific" quantities encourage its implementation as an additional "module" to standard gene-finders (e.g., GeneMark) improving their prediction rates (for a related work see: Almirantis and Nikolaou, 2004).

References

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

Akashi H, Kliman RM, Eyre-Walker A (1998) Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Genetica 102-103:49–60

Almirantis Y, Nikolaou C (2004) Multi-criterial coding sequence prediction. Combination of GeneMark with two novel, coding-character specific quantities. Comput Biol Med, in press

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G (1993) The isochore organization of the human genome and its evolutionary history—a review. Gene 135:57–66

Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. Proc National Academy of Sciences of the United States of America 83:5155–5159

Brendel V, Beckmann JS, Trifonov EN (1986) Linguistics of nucleotide sequences: morphology and comparison of vocabularies. J Biomol Struct Dyn 4:11–21

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1358–1362

Crick FH, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7:389–397

Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. A: Emergence of the hypercycle. Naturwissenschaften 64:541–565

Fickett JW (1996) Finding genes by computer: The state of the art. Trends Genet 12:316–it 320

Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20:6441–6450

Guigó R (1999) DNA composition, codon usage and exon prediction. In: Bishop Jm (ed) Genetic databases. Academic Press, New York

Gutiérrez G, Oliver JL, Marín A (1993) Dinucleotides and G+C Content in human genes: opposite behavior of GpG, GpC, and TpC at II-III codon positions and in introns. J Mol Evol 37:131–136

Gutiérrez G, Oliver J, Marin A (1994) On the orgin of the periodicity of three in protein coding DNA sequences. J theor Biol 167:413–414

Hanai R, Wada A (1989) Novel third-letter bias in Escherichia coli codons revealed by rigorous treatment of coding constraints. J Mol Biol 207:655–606

Hao BL (2000a) Fractals from genomes. Mod Phys Lett B 14:871–875

Hao BL (2000b) Fractals from genomes—Exact solutions of a biology-inspired problem. Physica A 282:225–246

Herzel H, Grosse I (1995) Measuring correlations in symbol sequences. Physica A 216:518–542

Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Res 18:2163–2170

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290

Karlin S, Ladunga I (1994) Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA 91:12832–12836

Karlin S, Mrazek J (1997) Compositional difference within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227–10232

Karlin S, Ladunga I, Blaisdell BE (1994) Heterogeneity of genomes: Measures and values. Proc Natl Acad Sci USA 91:12837–12844

Kliman RM, Eyre-Walker A (1998) Patterns of base composition within the genes of *Drosophila melanogaster*. J Mol Evol 46:534–541

Kraemer E, Wang J, Guo J, Hopkins S, Arnold J. (2001) An analysis of gene-finding programs for Neurospora crassa. Bioinformatics. 17:901–912

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 26:1107–1115

Makhoul CH, Trifonov EN. (2002) Distribution of rare triplets along mRNA and their relation to protein folding. J Biomol Struct Dyn 20:413–420

Marais G, Mouchiroud D, Duret L (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 98:5688–5692

Nikolaou C, Almirantis Y (2002) A study of the middle-scale nucleotide clustering in DNA sequences of various origin and functionality, by means of a method based on a modified standard deviation. J Theor Biol 217:479–492

Nikolaou C, Almirantis Y (2003) Mutually symmetric and complementary triplets: differences in their use distinguish systematically between coding and noncoding genomic sequences. J Theor Biol 223:477–487

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394

Nussinov R (1981) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 149:125–131

Rogic S, Mackworth AK, Ouellette FB (2001) Evaluation of gene-finding programs on mammalian sequences. Genome Res 11:817–832

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Shepherd JC (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Shepherd JC (1990) Ancient patterns in nucleic acid sequences. Methods Enzymol 183:180–192

Stuckle EE, Emmrich C, Grob U, Nielsen PJ (1990) Statistical analysis of nucleotide sequences. Nucleic Acids Res 18:6641–6647

Stuckle EE, Nielsen PJ, Grob U (1992) Probability of occurrence of specific oligomers. J Theor Biol 159:299–306

Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy , R (1997) Prediction of probable genes by fourier analysis of genomic sequences. Comp Appl in Biosci 13:263–270

Urrutia AO, Hurst LD (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159:1191–1199

Wong JT Cedergren R (1986) Natural selection versus primitive gene structure as determinant of codon usage. Eur J Biochem 159:175–180