Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 304 (2003) 183–192

GENE
AN INTERNATIONAL JOURNAL ON
GENES AND GENOMES

www.elsevier.com/locate/gene

Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*

Emanuele Bultrini[a], Elisabetta Pizzi[a,*], Paolo Del Giudice[b], Clara Frontali[a]

[a]*Laboratorio di Biologia Cellulare, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy*
[b]*Laboratorio di Fisica, Istituto Superiore di Sanità, Rome, Italy*

Received 4 July 2002; received in revised form 15 November 2002; accepted 4 December 2002
Received by G. Pesole

Abstract

Overall compositional properties at the level of bases, dinucleotides and longer oligos characterize genomes of different species. In *Caenorhabditis elegans*, using recurrence analysis, we recognized the existence of a long-range correlation in the oligonucleotide usage of introns and intergenic regions. Through correlation analysis, this is confirmed here to be a genome-wide property of *C. elegans* non-coding portions. We then investigate the possibility of extracting a typical vocabulary through statistical analysis of experimentally confirmed introns of sufficient length (> 1 kb), deprived of known splice signals, the focus being on distributed lexical features rather than on localized motifs. Lexical preferences typical of introns could be exposed using principal component analysis of pentanucleotide frequency distributions, both in *C. elegans* and in *Drosophila melanogaster*. In either species, the introns' pentamer preferences are largely shared by intergenic tracts. The pentamer vocabularies extracted for the two species exhibit interesting symmetry properties and overlap in part. A more extensive investigation of the interspecies relationship at the level of oligonucleotide preferences in non-coding regions, not related by sequence similarity, might form the basis of new approaches for the study of the evolutionary behaviour of these regions.

Keywords: Introns; *Caenorhabditis elegans*; *Drosophila melanogaster*; Linguistic properties

1. Introduction

It is a well known fact that, in eukaryotic genomes, oligonucleotide preferences differ between exons and introns. As early as 1986, Beckmann et al. (1986), by applying a linguistic approach (Brendel et al., 1986), revealed non-random patterns specific for introns. In the same year, Claverie and Bougueleret (1986), using the sequence data available at that time, compiled separate tables for the frequencies of overlapping oligonucleotides (1–9 bp long, shifted by 1 bp) in intron and exon sequences, and successfully constructed discriminant profiles on the basis of differences in pentamer or hexamer statistical distributions. In commenting on their results, they suggest that "intron and exon sequences have a distinctive statistical 'colour' or 'taste' irrespective of the underlying translation reading frame".

In-phase hexamer usage (Claverie et al., 1990; Fickett and Tung, 1992) is presently part of the standard toolbox for 'ab initio' gene-predicting algorithms. This protein-coding measure depends to some extent on the G + C content of the analyzed sequences (Guigò and Fickett, 1995). Indeed, corrections to mitigate the dependence on G + C content were introduced into some of these programs (e.g. GRAIL, GeneParser).

Independently of its performance as a predictor, the intrinsic bias in oligonucleotide usage observed by Claverie and Bougueleret (1986) – if not trivially due to systematic differences in base composition – is interesting *per se*. The question arises whether the observed bias can be attributed to just a few key patterns or reflects a more distributed feature. Lim and Burge (2001), in their analysis of the amount of information required for accurate intron recognition, focus on short (typically less than 100 bp) introns in five completely sequenced genomes (*Saccharomyces cere-*

Abbreviations: PCA, principal component analysis; PC1, first principal component; PC2, second principal component; ORF, open reading frame.

* Corresponding author. Tel.: +39-06-49902226; fax: +39-06-49387143.

E-mail address: epizzi@iss.it (E. Pizzi).

0378-1119/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0378-1119(02)01206-4

visiae, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Homo sapiens*). They find that – especially in some species – intron detection is not warranted by localized signals (the well known 5′, 3′ and branch motifs) alone, but improves when including a score for pentamer composition. They suggest that the existence of an intron vs. exon bias in pentamer preferences might reflect the presence of intronic splicing enhancers.

Previous evidence obtained from recurrence analysis of extended genomic regions (Frontali and Pizzi, 1999), on the other hand, suggests that introns share a set of preferred oligos with intergenic regions. In *C. elegans* this kind of cross-correlation between introns and intergenic regions was shown to extend over several megabases. However, no clue was offered by that approach as to the definition of the underlying vocabulary. We thus felt it was necessary to explore different ways to unravel the basis of the observed recurrence properties.

In the present paper we first confirm and extend the recurrence results by means of correlation analysis, using whole-genome data (complete sets of experimentally confirmed introns and exons) for *C. elegans*. In order to capture distributed lexical features which might characterize the introns' oligonucleotide usage, we then focus on the longest members of the intron set (>1 kb, artificially deprived of known 5′, 3′ and branch signals) and apply principal component analysis (PCA) to find the combinations of pentamers that yield the highest contribution to the separation between actual and randomized intron sequences and between intron and exon sequences.

The extracted pentamer list (the 'vocabulary') is then used to analyze coding and non-coding genomic portions in terms of their local content in vocabulary words. From this point of view, regions which – according to current annotation – are intergenic appear similar to introns. The same is true for *D. melanogaster*, whose similarly-extracted vocabulary partially overlaps that of *C. elegans*.

Finally, interesting self-complementarity properties of the extracted vocabularies emerge, which might usefully complement previous observations concerning single-strand symmetry (Prabhu, 1993; Baisnée et al., 2002).

2. Methods

2.1. Data source

Intron and exon sequences used in this analysis were extracted from the Exon-Intron Database (EID; Saxonov et al., 2000) available at http://golgi.harvard.edu/gilbert/eid, based on GenBank release 115. Only introns and exons experimentally confirmed through mRNA matching were considered. These are: 3296 introns (for a total of 1198 kb) and 3907 exons (895 kb in total) for *C. elegans*; 1409 introns (436 kb in total) and 2078 exons (741 kb in total) for *D. melanogaster*.

By eliminating from whole-chromosome data all the ORFs present in current annotation (ftp://ftp.ncbi.nih.gov/genbank/genomes), long sequences were obtained which are presumably composed only of intergenic tracts. This was done for *C. elegans* chromosome I (resulting in a sequence, also indicated as 'intergenic I', of 8.5 Mb out of 16.2 Mb) and chromosome II (intergenic II, 9.9 Mb out of 17.0 Mb). The same was done for *D. melanogaster* scaffolds AE002602 from the left arm of chromosome 3 (intergenic I, 10.3 Mb out of 15.1 Mb) and AE002787 from the right arm of chromosome 2 (intergenic II, 8.8 Mb out of 13.7 Mb).

2.2. Correlation analysis

In order to obtain a genome-wide analysis for *C. elegans*, the following strategy was applied. Experimentally confirmed introns or exons were joined head to tail (5′–3′) to construct an intron and an exon supersequence, to be used along with the intergenic supersequences resulting from the procedure applied to chromosomes I and II (see above).

Supersequences were subdivided into non-overlapping 1 kb windows. For each window, the frequencies of overlapping pentamers (shifted by one nucleotide) were recorded. The same was done after randomly shuffling the bases in each window. Pearson correlation coefficients were then calculated between pentamer frequency distributions for all pairs of windows either from the same or from different supersequences. Results were condensed in the form of histograms representing the distribution of the total population of correlation values between the following types of regions: intron vs. intron; intron vs. intergenic; exon vs. exon; exon vs. intergenic; intron vs. exon. In each case the same was done vs. shuffled versions.

2.3. PCA

PCA was performed using MATLAB (version 12.0) statistical toolbox on vectors having as elements the 1024 observed pentamer frequencies. Loadings, i.e. the correlation coefficients between each principal component and each of the original 1024 variables, express how much each pentamer contributes to a given principal component.

Among the experimentally confirmed intron and exon sequences extracted from complete genome data as described above, only those longer than 1 kb were considered for statistical reasons (shorter sequences would yield too many empty bins in pentamer distributions). These were: 256 introns and 67 exons for *C. elegans*; 87 introns and 37 exons for *D. melanogaster*. Motifs corresponding to 5′, 3′ splice and branch signals were then eliminated from each intron by cutting the first (5′) 8 bp and the last (3′) 50 bp ('shortened' introns). The sets to which PCA was applied are, for either organism: the set of shortened introns; the set obtained by randomizing each of these shortened introns three times and averaging the resulting frequency histo-

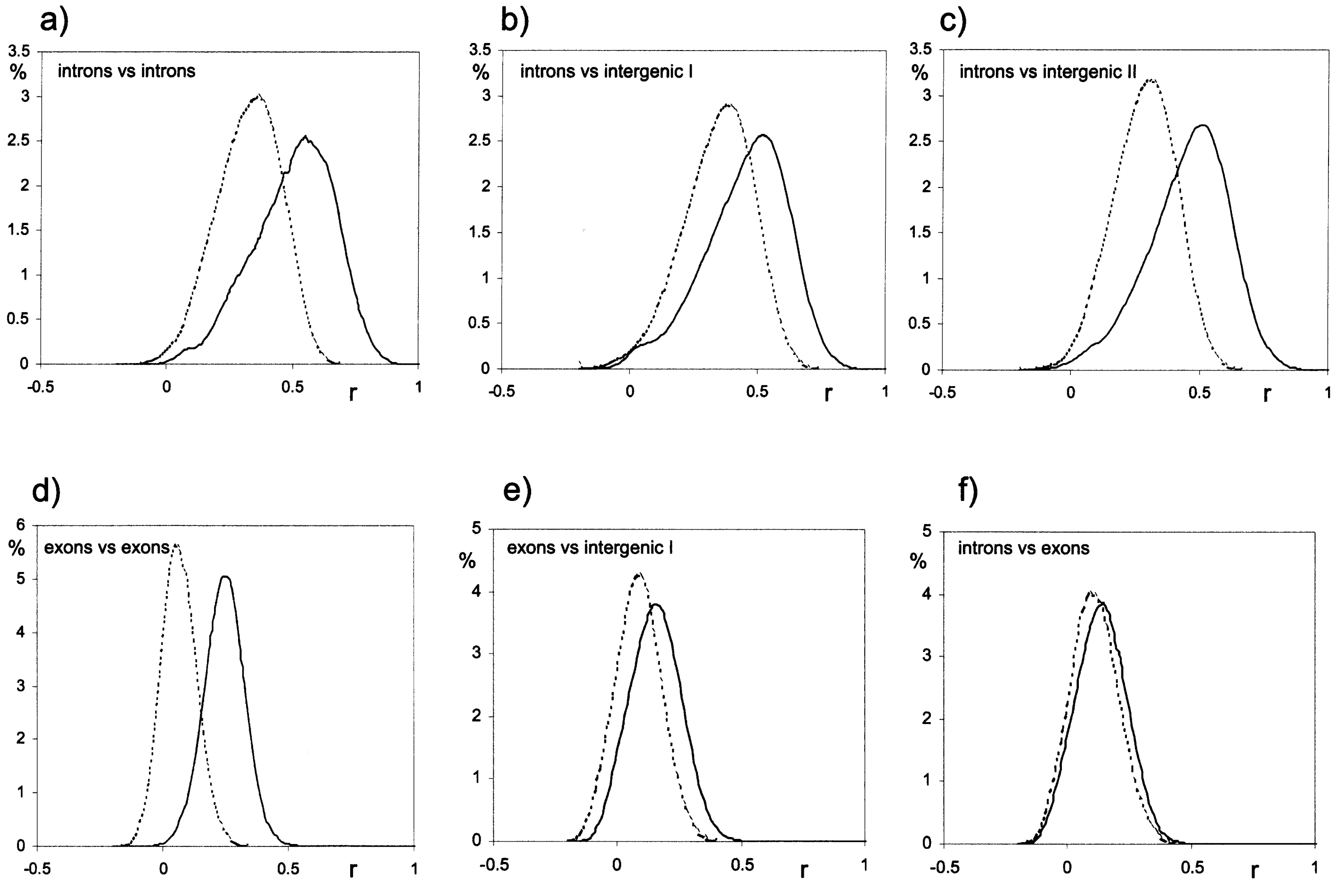

Fig. 1. Histograms representing populations of Pearson correlation coefficients, *r*, between pentamer frequency distributions for all possible pairs of non-overlapping 1 kb windows cut along intron, exon and intergenic *C. elegans* supersequences. (a) Intron windows against themselves and against their randomized versions; (b) intron windows against intergenic I (real and randomized) windows; (c) same as (b) for chromosome II; (d) exon windows against themselves and against their randomized versions; (e) exon windows against intergenic I (real and randomized) windows; (f) intron windows against real and randomized exon windows. Comparisons involving randomized data are given as dotted curves.

grams; the set of exons longer than 1 kb. Typical vocabularies were constructed as described in Section 3.

Tetramer bias on pentamer intron composition was assessed for *C. elegans* by estimating for each intron sequence: (a) pentamer occurrences expected on the basis of the observed tetramer occurrences ($[A_1A_2A_3A_4A_5] = [A_1A_2A_3A_4][A_2A_3A_4A_5]/[A_2A_3A_4]$); (b) pentamer occurrences observed after shuffling the original sequences so as to preserve tetramer composition using the Shufflet software (Coward, 1999). In either case, estimated pentamer occurrences, averaged over the entire set, were compared with those observed in the original set. Differences larger than three standard deviations were considered significant.

3. Results

3.1. Correlation analysis

Results obtained through recurrence analysis (Frontali and Pizzi, 1999) indicate the existence of a long-range correlation between pentamer frequency distributions in non-coding portions (introns and intergenic regions) of *C. elegans* chromosome III. In order to test whether this is a genome-wide property, we first constructed an intron supersequence by joining head to tail the 3296 experimentally confirmed introns that can be extracted from the complete *C. elegans* genomic data. By similarly joining the 3907 experimentally confirmed exons, an exon supersequence was constructed. Supersequences containing ordered collections of intergenic tracts were obtained as described in Section 2 for *C. elegans* chromosomes I and II.

We next partitioned the above supersequences into non-overlapping windows 1 kb long, and compared windows, either within or between supersequences, by pairwise calculating the Pearson correlation coefficients between their pentamer frequency distributions.

Results are reported in Fig. 1. Histograms show the distribution of correlation coefficients between all window pairs' intra- and inter-supersequences. The corresponding distributions involving randomized data (dotted curves) are also reported, so as to reveal whether systematic differences

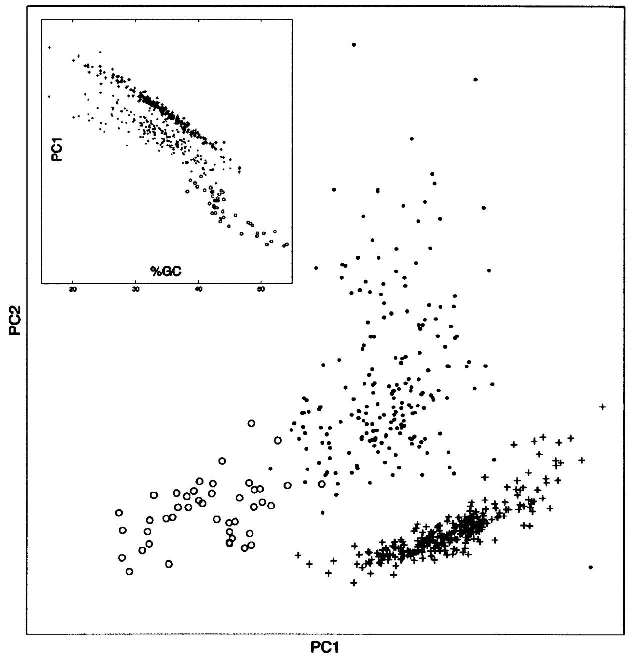

Fig. 2. PCA. Pentanucleotide frequency distributions for the experimentally-confirmed 256 introns and 67 exons longer than 1 kb from *C. elegans*, and for the set of randomized introns (see Section 2 for the elimination of splice signals from introns) were pooled and subjected to 'blind' PCA. The figure shows a scatter plot in the plane of the first two principal components (PC1 and PC2). Labels (full circles: introns; open circles: exons; crosses: randomized introns) were added *a posteriori*. Inset: PC1 values from the main figure are plotted against the G + C content of the corresponding sequences.

in base composition can account for the observed correlation effects.

Self-correlation within the intron supersequence (Fig. 1a) reveals a bias in pentamer preferences (median correlation coefficient 0.52) above the effect of base composition alone (dotted distribution). Similar results are obtained for intron vs. intergenic correlations (Fig. 1b,c), indicating that non-coding portions do exhibit a genome-wide similarity in pentamer composition beyond random expectation.

As it might have been anticipated, exons do not show any significant self- (Fig. 1d) or cross-correlation (Fig. 1e,f). In the two latter cases actual or shuffled sequences give almost superimposable distributions, suggesting that a systematic bias in base composition accounts for the lack of correlation. The small pentamer bias (median correlation coefficient 0.25) observed when pairwise comparing exon windows (Fig. 1d) is most probably the consequence of preferences in codon usage.

3.2. *Caenorhabditis elegans introns' vocabulary*

In order to find out whether the correlation effects observed at the genome scale in *C. elegans* non-coding portions (introns and intergenic regions) involve co-variation of a limited number of words, we compared oligonucleotide frequency distributions for the set of experimental introns, for the set obtained by randomly shuffling the bases of each intron, and for the set of experimental exons.

Known splice signals obviously affect such distributions in the case of introns. Being interested in distributed linguistic features rather than in localized signals, we restricted our analysis to those introns that are sufficiently long to provide significant statistics even when $5'$, $3'$ and branch motifs are eliminated (see Section 2). This led us to select the 256 experimentally confirmed *C. elegans* introns that are longer than 1 kb (the longest sequence in the intron set being 15 kb).

Observed pentamer distributions were considered as vectors whose components are the 1024 observed pentamer frequencies. These frequencies do not represent independent variables, not only because we are considering overlapping oligos, but also because of their possible combination into longer characteristic words. The symmetry properties discussed below also affect frequency interdependence in the case of reverse complementary pentamers.

Since the dimensionality of the space is significantly larger than the number of data points, usual classification methods (such as linear discriminant analysis) would be underdetermined, and a dimensional reduction of the problem should precede the analysis. We therefore carried out PCA so as to reduce the dimensionality of the system and to simultaneously highlight the presence of any intrinsic structure of the data.

When applied to real and randomized introns, PCA turned out to provide a good distinction of the two data sets already in a scatter plot of the first two principal components, the dispersion of each cloud being fairly isotropic and well below the distance between the two gravity centres. When the *C. elegans* experimentally confirmed exons longer than 1 kb were added to the pool of unlabeled objects to be analyzed by PCA, three clouds became visible in the scatter plot of the two new principal components (Fig. 2). Little or no separation was associated with the 3rd or 4th principal components. So, the first two principal components turned out *a posteriori* to provide an effective classification tool for *C. elegans* introns, exons and randomized introns.

As shown in the inset in Fig. 2, the first principal component, PC1, which yields good separation between exons and (real or randomized) introns, closely reflects the base composition (G + C content) of the sequences in the three sets. Conversely, PC2 is not significantly related to G + C content, as might have been expected given that it affords a good separation between real and randomized introns. Along the latter component, exons occupy a range that covers that of randomized introns and only partly overlaps that of real introns.

As a next step, we determined the loadings (as defined in Section 2) on PC2 for the 1024 original variables. Positive values, in the present case, correspond to oligos that are

Table 1
Introns' pentamer vocabularies

Caenorhabditis elegans		*Drosophila melanogaster*	
AATTT*	AAATT*	**AAAAT**	**ATTTT**
AAAAT*	**ATTTT***	AAAAG	CTTTT
GAAAA*	**TTTTC***	**GTTTT**	**AAAAC**
TTTTT*	AAAAA*	AAATG	CATTT
TTTTG	**CAAAA***	AAACA	TGTTT
TGAAA	TTTCA	AATGC	GCATT
GGAAA	TTTCC*	AACAA	TTGTT
CGAAA	TTTCG	**TTTTG**	**CAAAA**
GATTT*	AAATC*	**GAAAA**	**TTTTC**
TTGAA*	TTCAA	TTTGG	CCAAA
TTTGA	TCAAA	TTTGC	GCAAA
GTTTT*	**AAAAC***	TTGGC	GCCAA
		TTGCA	TGCAA
GAAAT*, **AGAAA***		**AGAAA**, AGCAA, **TTTTT**	
		TTTGT, **TGAAA**, TGGCA	
		GAAAT, GGCAA, GCAGC	
		GCAAC, CAAAT, CAATT	
		CAACA, CAGCA	

Pairs of reverse complementary pentamers (ranked according to PC2 loadings) are listed separately from unpaired ones. Asterisks mark pentamers significantly different from tetramer-based expectation. Pentamers appearing in both lists are in bold.

consistently over-represented in real vs. randomized intron sequences, while negative loadings are associated with oligos that are avoided in the real intron sequences. Table 1 lists the 26 pentamers that contribute to PC2 with a loading greater than or equal to +0.50. We checked that reducing the 1024 original variables to the subset of 26 pentamers does not significantly reduce the separation between real and randomized introns observed in Fig. 2. Hereafter we will refer to the list in Table 1 as the *C. elegans* introns' vocabulary.

The occurrence of the four bases in the list in Table 1 reflects – as one might expect – the A + T bias of the intron set, whose base composition is given in Table 2. However, clear preferences appear among pentamers containing the same bases in a different order: for instance, while AAATT and AATTT have PC2 loadings that are amongst the highest in the list, pentamers like ATATA and TATAT or ATTAA and TAATT are not over-represented with respect to random expectation.

To investigate whether the pentamer vocabulary results from preferences already existing at the level of tetramers, we compared observed pentamer occurrences (averaged over the entire intron set) with those expected 'a priori' on the basis of the average occurrences observed for tetramers (see Section 2). It could thus be shown that 17 out of the 26 vocabulary words (marked by asterisks in Table 1) differ significantly from tetramer-based expectation. This was confirmed by measuring average frequencies of pentamers in sets obtained by shuffling each intron so as to preserve tetramer composition (Shufflet analysis, see Section 2). Statistical limits due to the average size of the introns prevented us from investigating whether non-trivial preferences still appear for longer words, possibly composed of overlapping oligos from the extracted list.

Table 2
Base composition and skewness in *C. elegans* introns and exons (>1 kb)

	% A	% T	% G	% C	(A − T)/(A + T)
Introns	32.3	33.8	16.5	17.4	−0.023
Exons	30.5	25.2	23.4	20.9	0.095

3.3. Symmetry properties of the *C. elegans* introns' vocabulary

The extracted vocabulary is almost entirely composed of pairs of reverse complementary pentamers, reported in separate columns in Table 1. Is this a trivial consequence of a base composition conforming to what is generally known as the second Chargaff's parity rule (A = T and G = C along the same strand)?

Average base composition for the sets of experimental introns and exons utilized in PCA (Section 3.2) is reported in Table 2. In fact introns follow the parity rule much better than exons, where nucleotide preferences associated with defined codon positions contribute to strand asymmetry.

It is thus to be expected that sufficiently long sequences with the base composition of introns will exhibit almost perfect symmetry (Baisnée et al., 2002) even if randomly shuffled at the nucleotide level. It is, however, interesting to note that the symmetrical trend is already evident for individual introns – but not for their randomized counterparts – at a length scale at which statistics do not afford the asymptotic behaviour. This is shown in Fig. 3, reporting scatter plots for the frequencies of the two members of each of the 1024/2 = 512 reverse complementary pairs in 241 introns of length comprised between 1 and 15 kb or in their shuffled versions. (A few introns, individually exhibiting an AT skewness higher than 0.15 in absolute value, i.e. differing by more than two standard deviations from the average value reported in Table 2, were omitted in this analysis.)

The 12 pentamer pairs belonging to the extracted vocabulary make the largest contribution to this trend: in effect, when they are excluded (Fig. 3c), the remaining 500 pairs yield data points that remain confined near the origin for all of the 241 introns.

One can thus conclude that – notwithstanding the large sampling errors – individual *C. elegans* intron sequences are far more symmetrical at the pentamer level than could have been expected on the basis of their nucleotide composition.

3.4. Genomic distribution of the introns' vocabulary

The vocabulary defined in the preceding sections identifies pentamers that occur in *C. elegans* introns above

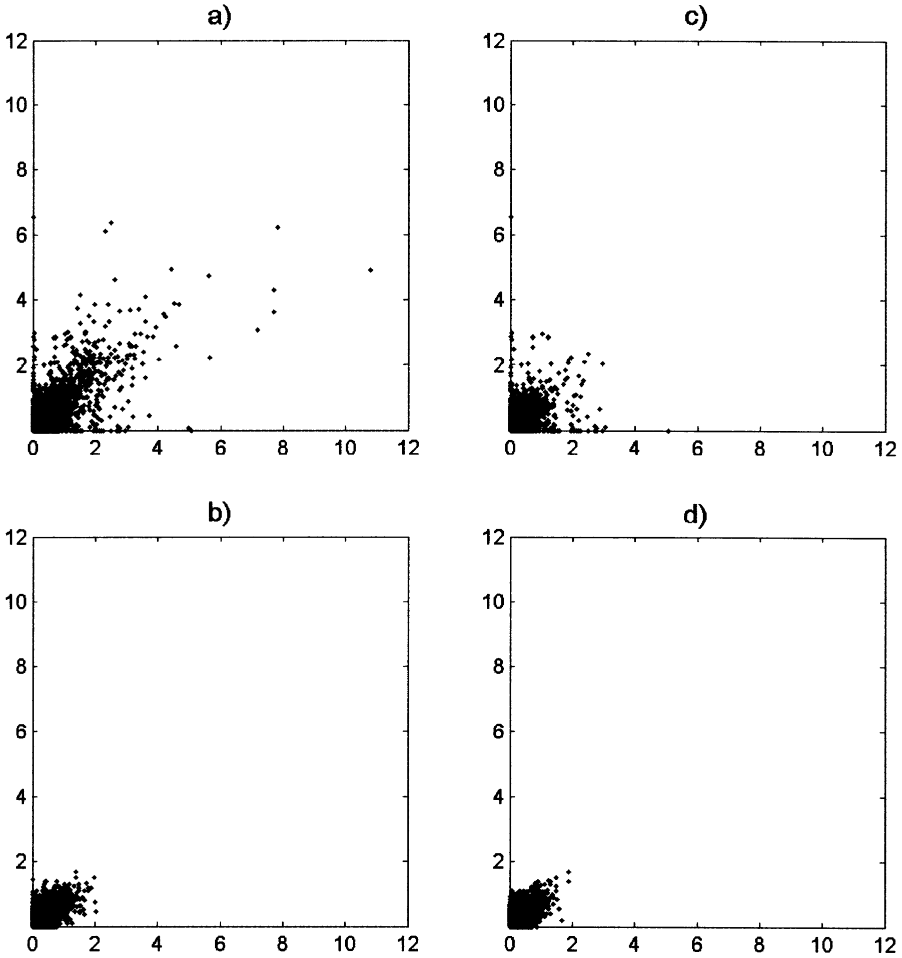

Fig. 3. Symmetry scatter plots for *C. elegans* introns longer than 1 kb and their randomized counterparts: (a) the 512 pairs of reverse complementary pentamers are indicated by dots having as co-ordinates the frequencies (expressed as percent values) of either member of the pair in each of the 241 introns examined; (b) same as (a) for the 241 randomized intron sequences; (c) same as (a) after exclusion of the 12 pairs appearing in *C. elegans* introns' vocabulary; (d) same as (c) for the randomized intron sequences.

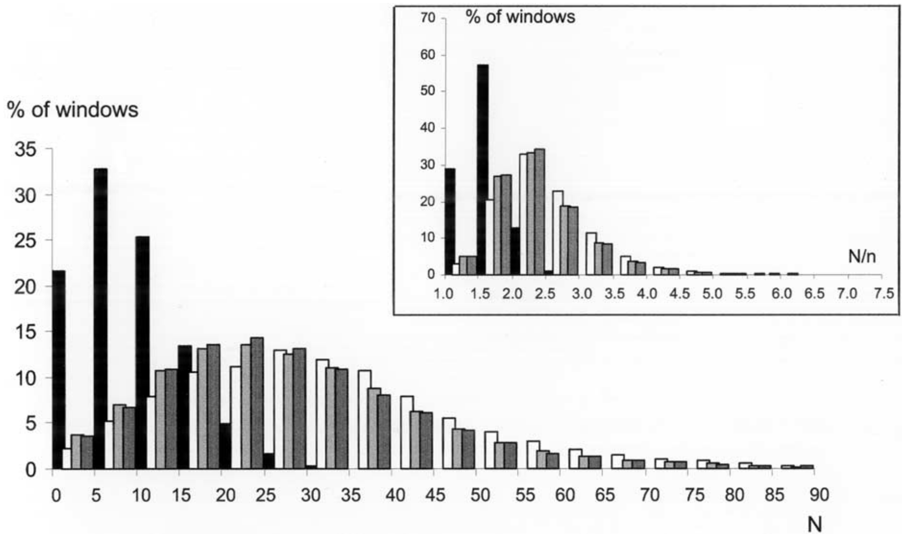

Fig. 4. Vocabulary usage in different regions of the *C. elegans* genome. Distributions according to the content, *N*, in vocabulary words are given for the populations of 200 bp non-overlapping windows from the intron (white), intergenic I (light grey), intergenic II (dark grey) and exon (black) supersequences. In the inset the same sets of windows are analyzed in terms of the ratio *N*/*n*, *n* being the number of different vocabulary words present in the window.

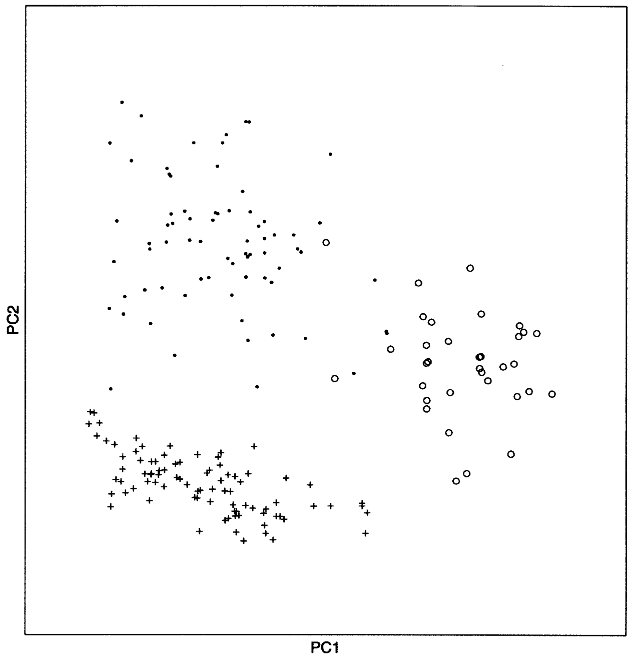

Fig. 5. PCA for the experimentally confirmed 87 introns and 37 exons longer than 1 kb from *D. melanogaster*, and for the set of randomized introns. Symbols are as in Fig. 2.

random expectation based on nucleotide composition. In order to test such a vocabulary's prevalence in other genomic portions we made use of the intron, exon and intergenic supersequences already used in correlation analysis (Section 3.1). Supersequences were subdivided into non-overlapping 200 bp windows, and for each window the number of vocabulary words, N, and the number, n, of different vocabulary words used were recorded. Histograms showing the percentual distribution of windows according to their content in vocabulary words, N, are shown in Fig. 4. In the inset the same is done with respect to the ratio N/n, expressing the average repetitivity of the vocabulary words present in a window.

Intergenic regions closely follow the behaviour of introns in the local use of the vocabulary. About 55% of the intergenic 200 bp windows contain at least 25 vocabulary words, to be compared with 66% of the intron windows of the same length. Only 7% of the exon windows contain 25 or more vocabulary words.

The widespread use in intergenic regions of the pentamer vocabulary preferred by introns is thus confirmed. From a study of the length distributions of intergenic segments formed by contiguous windows of similar word content, it appears that this intron-like behaviour spreads over multiple length scales (at least above the 200 bp coarse-graining we used in order to avoid exceedingly high sampling errors in word counts).

Regional profiles can be obtained by reporting counts (N) of vocabulary words in a window (e.g. 200 bp long) sliding along limited genomic portions: in *C. elegans* these profiles reveal frequent intergenic peaks separated by short valleys of low N (data not shown).

3.5. *Drosophila melanogaster* introns' vocabulary

We then proceeded to investigate whether the observed properties are peculiar for *C. elegans* introns or whether they also characterize introns from *D. melanogaster*. Fig. 5 presents the PC1/PC2 plot for introns, randomized introns and exons in this latter organism. Again, the first principal component essentially reflects a systematic difference in G + C content between exons and introns, while actual and randomized introns are best separated along the second principal component.

The approximation utilized in defining the introns'

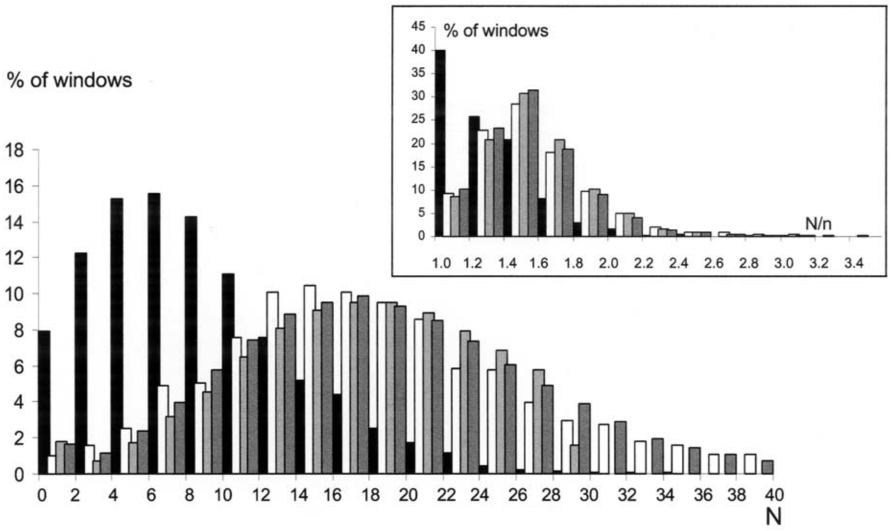

Fig. 6. Vocabulary usage in different regions of the *D. melanogaster* genome. Distributions according to the content, N, in vocabulary words are given for the populations of 200 bp non-overlapping windows from the intron (white), intergenic I (light grey), intergenic II (dark grey) and exon (black) supersequences. Inset as in Fig. 4.

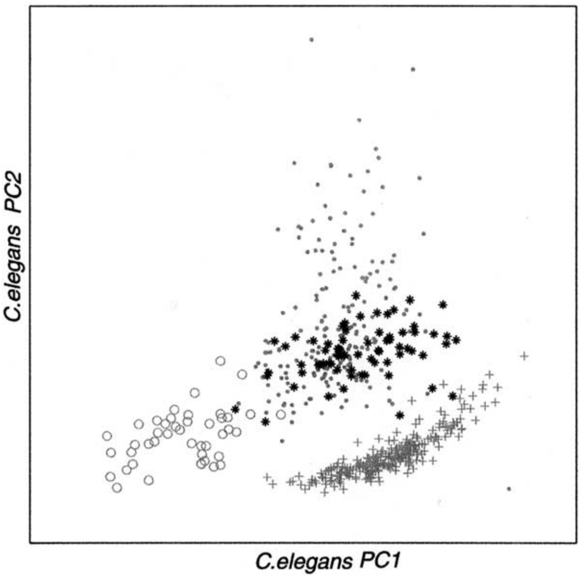

Fig. 7. Projection (see text) of the *D. melanogaster* intron set (black stars) onto the *C. elegans* PC1/PC2 plane of Fig. 2, reproduced in grey.

vocabulary for *C. elegans* (i.e. to neglect all pentamers contributing to the second principal component with loadings lower than 0.50) corresponds in *D. melanogaster* to the selection of the 40 top-ranking pentamers. These are listed in Table 1, defining the *D. melanogaster* introns' vocabulary. Though less marked than in *C. elegans*, reverse complementarity still holds for 26 of the 40 vocabulary words.

The extracted vocabulary was then tested on intron, exon and intergenic *D. melanogaster* supersequences as already done for the nematode. Fig. 6 is in fact the equivalent of Fig. 4. Although the difference between coding and non-coding sequences is smaller here than in the case of *C. elegans*, the similarity in pentamer usage between introns and intergenic regions appears to be confirmed also in the case of *D. melanogaster*.

The two vocabularies share the words marked by bold characters in Table 1. A different way to show that the introns in the two species share several pentanucleotide preferences is illustrated in Fig. 7, which shows how *D. melanogaster* introns are distributed in the PC1 vs. PC2 plot of Fig. 2 (reproduced in grey). To construct this figure, the matrix transforming the 1024 pentamer frequencies into principal components for *C. elegans* was applied to the pentamer frequency vectors representing the introns of the *D. melanogaster* set. This procedure allowed us to 'project' the fly's introns onto the bidimensional PC1/PC2 space that provides separation between *C. elegans* introns, randomized introns and exons. In this presentation, the fly's introns (black stars) turn out to fall within the *C. elegans* introns' distribution.

Comparison with other eukaryotic species is presently the object of a more detailed study, which must also take into account the isochore organization of the genomes of warm-blooded vertebrates (Bernardi et al., 1985; Bernardi, 1995).

4. Discussion

Rather than intron prediction, which must take into account various factors (Lim and Burge, 2001), the aim of the present work is to extract and analyze the lexical features that, apart from the known splice signals, characterize introns. The results of our statistical analyses, performed on sufficiently long (>1 kb) and experimentally confirmed introns from either *C. elegans* or *D. melanogaster*, show that it is possible to extract lists of pentamers significantly over-represented in the introns of either species with respect to random expectations. We refer to these lists as the introns' vocabularies.

The approach we followed to extract typical vocabularies involves PCA of frequency distributions in overlapping pentamers (read in the direction in which they are transcribed). This approach identifies co-variance properties that distinguish introns from their randomized counterparts and from exons. While introns and exons, both in *C. elegans* and *D. melanogaster*, are best separated along the first principal component (which essentially reflects a difference in nucleotide composition) the linear combination of pentamers forming the second principal component is the one that best separates real from shuffled introns. For these two organisms, therefore, it is possible to achieve an effective reduction from 1024 variables to just two principal components.

Comparison with observed tetramer frequency distributions confirms that the extracted pentamer set is not explained by the presence of shorter patterns or motifs, so that subtracting the contribution of shorter oligos was not deemed essential. On the other hand, the choice to consider five-letter words – dictated by statistical reasons – provides hints for the existence of longer words resulting from the superposition of those identified as members of a typical vocabulary.

In principle there is no reason to believe that words of any fixed length should emerge as a prominent statistical feature. This is clearly stated by Bussemaker et al. (2000) in their presentation of a statistical method addressing the difficult task of building up a dictionary of non-overlapping, possibly unique words or motifs occurring in functionally related genomic regions much more frequently than expected from the juxtaposition of shorter words of various lengths.

The PC2-extracted vocabulary is not directly comparable to the list of pentamers sorted by Lim and Burge (2001) on the basis of their contribution to relative entropy of intron vs. exon oligomer composition, through a procedure that assumes a homogeneous 4th order Markov chain model to be valid for both coding and non-coding sequences. Following our line of reasoning, one would expect that the top ten 'intron-biased' pentamers reported by Lim and Burge for *D. melanogaster* would score high in terms of loading on PC1, and low on PC2. We verified that this was, in fact, the case. Furthermore, it should be stressed that – in their search for the minimal elements warranting correct intron recognition – Lim and Burge (2001) focus on short introns (less than 60 and 80 bp in *C. elegans* and *D. melanogaster*, respectively). This choice, perfectly suited to their purpose, is not adequate for investigating properties that prevail in intron portions possibly not involved in the splicing process, and that instead may be considered as free to drift, or even as junk DNA. In a sense, our work is complementary to that of Lim and Burge and tries to answer the following questions: irrespective of the elements known to be involved in intron recognition, do introns obey 'linguistic constraints'? And, if so, are these constraints specific for a given genome?

In the two organisms analyzed in the present work, the answer to the first question is positive. Consistent properties emerge from the study of experimental introns from different genomic locations, and a 'genome effect' at the level of oligonucleotide preferences in non-coding portions (introns and intergenic regions) appears to be superimposed onto nucleotide preferences (see also Gentles and Karlin, 2001).

In fact the preference for the vocabulary that characterizes introns appears to be widely shared by intergenic regions. This finding might be surprising, since it is reasonable to consider these regions as mosaics of tracts involved in a variety of functions (or even not functional at all). Regulatory functions – and in general genomic tracts deputed to specific interactions – might be expected to be characterized by a minimal use of the introns' vocabulary, since it is hard to think that the constraints implicit in the 'intron-like' behaviour would be compatible with the encoding of specific functions. So, the fact that – at least in *C. elegans* and *D. melanogaster* – a large fraction of intergenic DNA exhibits this behaviour most probably masks short, possibly functional elements that adopt specific preferences.

Concerning the genome specificity of the observed lexical properties, it is interesting to note that the vocabularies prevailing in *C. elegans* and *D. melanogaster* non-coding portions share a subset of pentamers. It thus appears that the evolutionary distance between the two organisms – albeit sufficient to reduce sequence conservation in these portions to undetectable levels – did not completely alter their linguistic preferences. This suggests that it might be possible to exploit partial conservation of lexical features, such as the introns' vocabulary, as a means to characterize relatedness between non-coding portions from different organisms, not related by sequence similarity, thus gaining knowledge on the evolutionary behaviour of introns and intron-like intergenic tracts. A more extensive study of the inter-species relationship concerning similarity in oligonucleotide preferences is necessary in order to confirm this guess.

Finally, a very interesting feature of the *C. elegans* intron vocabulary is its being almost entirely composed of pairs of reverse complementary oligos. The same property, though less marked, is present also in *D. melanogaster*. As shown by Prabhu (1993) for a wide range of organisms, reverse complementary oligos exhibit similar numbers of occurrences along sufficiently extended single-stranded regions that encompass genes in both direct and reverse orientation, as well as variable amounts of non-coding DNA. With a few exceptions, this observation is confirmed for an even wider set of prokaryotic and eukaryotic genomes by Baisnée et al. (2002), who developed methods to characterize symmetry at different orders (from the single nucleotide level up to 9 bp oligomers). In discussing the origin of the ubiquitous reverse-complement symmetry observed over sufficiently long single strands, the latter Authors stress the existence of higher order symmetry constraints whose contributions (residual symmetry) can be distinguished from lower order effects.

It is worthwhile noting that this symmetry is not present in exons aligned in one and the same direction, that generally exhibit a pronounced skewness effect. ($(A - T)/(A + T)$ in the complete sets of experimentally confirmed exons is of the order of 0.1 both for *C. elegans* and *D. melanogaster*.) On the contrary, notwithstanding the large sampling errors, a symmetrical trend is apparent on a scale of a few kilobases in individual *C. elegans* introns. This short-range property of introns is not simply due to their symmetrical base composition, since it is drastically reduced in randomized introns. Rather it results from the preferred use of reverse complementary oligomers, and in particular of the 12 pairs of pentamers present in the *C. elegans* introns' vocabulary. It would be tempting to link the above symmetry properties of introns to formation of stem-and-loop structures (reviewed in Forsdyke and Mortimer, 2000), but this would be an unjustified inference without a careful determination of the actual length of palindromes and of their mutual distances.

Taken together, the evidence presented suggests that, in the two studied species, the largely symmetrical vocabulary characterizing longer introns is locally shared by a large fraction of intergenic DNA.

Acknowledgements

Dr Alessandro Giuliani (Istituto Superiore di Sanità, Rome) provided us with constant and valuable advice. Constructive criticisms by Dr Wentian Li (North Shore University Hospital, Manhasset, NY) were essential in shaping the work. Both contributions are here gratefully acknowledged.

References

Baisnée, P.-F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? Bioinformatics 18, 1021–1033.

Beckmann, J.S., Brendel, V., Trifonov, E.N., 1986. Intervening sequences exhibit distinct vocabulary. J. Biomol. Struct. Dyn. 4, 391–400.

Bernardi, G., 1995. The human genome: organisation and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, J., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Brendel, V., Beckmann, J.S., Trifonov, E.N., 1986. Linguistic of nucleotide sequences: morphology and comparison of vocabularies. J. Biomol. Struct. Dyn. 4, 11–21.

Bussemaker, H.J., Li, H., Siggia, E.D., 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc. Natl. Acad. Sci. USA 97, 10096–10100.

Claverie, J.-M., Bougueleret, L., 1986. Heuristic informational analysis of sequences. Nucleic Acids Res. 14, 179–196.

Claverie, J.-M., Sauvaget, I., Bougueleret, L., 1990. K-tuple frequency analysis, from intron/exon discrimination to T-cell epitope mapping. Methods Enzymol. 183, 237–252.

Coward, E., 1999. Shufflet, shuffling sequences while conserving the k-let counts. Bioinformatics 15, 1058–1059.

Fickett, J.W., Tung, C.-S., 1992. Assessment of protein coding measures. Nucleic Acids Res. 20, 6441–6450.

Forsdyke, D.R., Mortimer, J.R., 2000. Chargaff's legacy. Gene 261, 127–137.

Frontali, C., Pizzi, E., 1999. Similarity in oligonucleotide usage in introns and intergenic regions contributes to long-range correlation in the C. elegans genome. Gene 232, 87–95.

Gentles, A.J., Karlin, S., 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res. 11, 540–546.

Guigò, R., Fickett, J.W., 1995. Distinctive sequence features in protein coding, genic non-coding and intergenic human DNA. J. Mol. Biol. 253, 51–60.

Lim, L.P., Burge, C.B., 2001. A computational analysis of sequence features involved in the recognition of short introns. Proc. Natl. Acad. Sci. USA 98, 11193–11198.

Prabhu, V.V., 1993. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 21, 2797–2800.

Saxonov, S., Daizadeh, I., Fedorov, A., Gilbert, W., 2000. An exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res. 28, 185–190.