

Introns resolve the conflict between base order-dependent stem-loop potential and the encoding of RNA or protein: further evidence from overlapping genes

I.H. Barrette, S. McKenna, D.R. Taylor, D.R. Forsdyke*

Department of Biochemistry, Queen's University, Kingston, Ontario, Canada K7L3N6

Received 10 December 2000; received in revised form 8 March 2001; accepted 5 April 2001

Received by D. Lilley

Abstract

Many eukaryotic genes are split into exons and introns, the latter being removed post-transcriptionally so that only exon sequences appear in cytoplasmic RNAs. Since introns appear in both protein-encoding RNAs and non-protein-coding RNAs, they interrupt genetic information *per se*, not just protein-encoding information. A DNA sequence has the potential to carry more than one type of genetic information, but different types may conflict. Thus, it has been proposed that introns arose because sequences were unable to contain concomitantly complete information for the encoding both of stem-loops and of cytoplasmic products (protein and/or RNA). Stem-loop potential is held to be selectively advantageous since it promotes the recombination-dependent correction of genetic errors. Stem-loop potential, the best local measure of which is base order-dependent stem-loop potential, tends to be less in exons than in introns. This is particularly evident in genes evolving rapidly under positive Darwinian selection, where the protein-encoding function is dominant. Evidence is now presented that the rare regions where genes overlap also impose excessive encoding demands so that the concomitant coding of base order-dependent stem-loop potential is decreased. Our results are consistent with the hypothesis that sequences with high stem-loop potential arose in the early 'RNA world'. Ancestors of modern genes would have entered this world when sequences (exons) encoding cytoplasmic products, were interspersed with sequences (introns) encoding selectively advantageous stem-loops. Purine-loading pressure would also have favoured intron formation. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Base order; Information conflict; Introns; Overlapping genes; Purine-loading; Recombination; Secondary structure; Stem-loop potential

1. Introduction

The sequence of a biological nucleic acid is the outcome of numerous, potentially conflicting, evolutionary pressures. These include the pressures to encode RNAs and proteins, to adopt a particular GC%, to purine-load mRNA-synonymous strands, and to form stem-loops (Forsdyke and Mortimer, 2000; Forsdyke, 2001a,b). One consequence of this seems to be that in many species the encoding of an RNA or protein cannot be achieved by one contiguous sequence of DNA bases (e.g. one open reading frame; ORF). Instead, the RNA or protein is encoded in segments ('exons'), which are interrupted by segments ('introns') which do not usually encode cytoplasmic products. The pressure to form stem-

loops, a pressure whose adaptive value may relate to meiotic synapsis, appears of particular importance in this respect.

Muller (1922) has suggested that the pairing of genes as parts of chromosomes undergoing meiotic synapsis, might provide clues on gene structure and replication. He noted that: "It is evident that the very same forces which cause the genes to grow should also cause like genes to attract each other, ... If the two phenomena are thus dependent on a common principle in the make-up of the gene, progress made in the study of one of them should help in the solution of the other". In 1954 he set his students an essay 'How does the Watson-Crick model account for synapsis?' (Carlson, 1981). Crick (1971) took up the challenge with his 'unpairing postulate' by which the two strands of the classical DNA duplex would unpair to allow a homology search. The unpaired single strands would form stem-loop structures, and this would facilitate synapsis and recombination (Doyle, 1978). Maternal and paternal chromosome homologs would mutually explore each other and test for 'self' DNA complementarity, using a loop-loop 'kissing'

Abbreviations: ORF, open reading frame; mRNA, messenger RNA; rRNA, ribosomal RNA

* Corresponding author. Tel.: +1-613-533-2980; fax: +1-613-533-2497; web site: <http://post.queensu.ca/~forsdyke/bioinfor.htm>.

E-mail address: forsdyke@post.queensu.ca (D.R. Forsdyke).

mechanism (Eguchi et al., 1991). If sufficient complementarity were found, then crossing over and recombination would occur. The main adaptive value of this would be to provide for the correction of errors in the individual homologs (Winge, 1917; Forsdyke, 1981, 1996b).

A potential conflict between protein-encoding capacity and nucleic acid secondary structure was recognized when the first sequences became available (Salser, 1970). However, this was considered the likely result of pressures operating at the mRNA level (Ball, 1973), a view held even by some recent commentators (Seffens and Digby, 1999). A role of genomic forces was suggested by the finding that stem-loop potential was genome-wide, affecting both genic and non-genic DNA (Forsdyke, 1995a–c; Heximer et al., 1996). It was proposed that the potential would conflict with a sequence's protein-encoding function to such an extent that open reading frames (ORFs) would have to be interrupted by introns in order to accommodate stem-loops. If so, the conflict would be most evident when the sequence was under strong selective pressure with respect to protein function. This was found to hold for genes evolving rapidly under positive Darwinian selection (Forsdyke, 1995b, 1996a). It should also hold for overlapping genes. Protein-encoding segments of genes which overlap protein-encoding segments of other genes should be more constrained than segments which do not overlap, and so should more powerfully exclude stem-loop potential. We here explore this hypothesis.

2. Methods

2.1. Sequences

Genes which overlap other genes were sought in databases using keywords such as 'overlap' and 'antisense'. Selection criteria were that overlaps should involve entire exons or ORFs, and that there should be comparable non-overlapping exons or ORFs in close proximity (i.e. in the same local genomic environment).

2.2. Secondary structure analysis

A natural sequence is but one member of a large set of possible sequences with the same base composition. Members of the set can be derived by repeatedly randomizing the natural sequence. Randomization (shuffling) destroys information present in the primary sequence (base order), without changing base composition or sequence length. Provided length is kept constant, an *average* characteristic of derived shuffled sequences should reflect base-composition alone (Bronson and Anderson, 1994). If the value of a quantifiable average characteristic is subtracted from the value of the corresponding characteristic of the natural sequence, one can determine if, and to what extent, there is a contribution of base-order to the characteristic in the natural sequence.

The characteristic studied here, stem-loop potential, was evaluated as described previously (Forsdyke, 1995d, 1998), using energy values for base stacking and loop destabilization assigned by Santalucia et al. (1996). In chemical thermodynamic terms, the formation of a helix in a stem-loop structure is favoured to the extent that free energy is released. A loss of free energy from molecules is expressed in negative kilocalories/mol. The greater this negative value, the more stable the resulting structure. Since extrusion tends to be an all-or-none phenomenon, stem-loops are more likely to be stably extruded from a region in which the average stem-loop potential has a high negative value, than from a region in which the average stem-loop potential has a low negative, or positive, value.

Minimum free energy values for the folding of a natural sequence (FONS values) are first determined for successive 200 nucleotide windows (moving in steps of 25 nucleotides). This is a function of both base composition and base order and measures the total stem-loop potential of a window. Then ten randomized sequences derived from each window are folded. The mean minimum free energy value for the ten sequences (the folding of randomized sequences mean value; FORS-M) provides a measure of the contribution of the base composition to the stem-loop potential of a window (the base composition-determined component of the stem-loop potential). Since FONS values are usually more negative than FORS-M values, differences between the two values (FONS less FORS-M) are usually negative. This difference (the folding of randomized sequence difference value; FORS-D) provides a measure of the contribution of base order alone to the stem-loop potential of a sequence window (the base order-determined component of the stem-loop potential).

It should be noted that in some previous studies (Forsdyke, 1995a–d; 1996a,b) the direction of the subtraction was reversed (FORS-M *less* FONS), so that FORS-D values were positive when the base order-dependent stem-loop potential was high. A rationale for the choice of 200 nt windows is given in Forsdyke, 1995d. To relate to the terminology of Le and Maizel (1989), it should be noted that $FONS = E$, $FORS-M = E_r$, and $FORS-D = \Delta E$.

2.3. Statistics

The argument of this paper rests on the extent to which the average of eight FORS-D difference values (penultimate column of Table 1) is significantly greater than zero. However, for each organism we have determined the probability that average FORS-D values differ between two types of segments (final column of Table 1). The FORS-D value for a window depends on the type and order of bases within the window. For our purpose, a window of 200 nt is defined as within an exon if its center overlaps the exon. Thus, there are as many potential windows as there are bases in the exon, and this number sets an upper limit on the number of possible samplings. Because the folding of

Table 1
Base order-dependent stem-loop potentials of gene segments which overlap other gene segments^a

Organism	Gene or genome segment	Accession number(s)	Overlap segments ^b		Non-overlap segments		FORS-D difference ^c	<i>P</i> ^d
			Exons/ORFs	FORS - D	Exons/ORFs	FORS - D		
T4 phage	30.1 to Ligase	X60109 X53848 X00039	30.3, ORF Y	-1.38 ± 0.64	30.6, 30.7, Ligase	-3.08 ± 0.26	1.7	0.004
G4 phage	Genome	V00657	A, D	1.28 ± 0.32	F, G, H	-2.28 ± 0.49	3.57	< 0.001
<i>Achlya klebsiana</i>	Glutamate dehydrogenase	U02505	Exon 10	-1.73 ± 0.35	Exons 3–9	-3.29 ± 0.94	1.55	0.077
<i>Drosophila melanogaster</i>	Fructose 1,6-bisphosphate aldolase	M98352	Exon 4A	1.24 ± 0.85	Exons 2, 3, 4C	-0.81 ± 0.56	2.04	0.066
<i>Rattus norvegicus</i>	Gonadotrophin releasing hormone	M31670	Exons 2, 3	3.10 ± 0.65	Exons 1, 4	0.68 ± 0.53	2.23	0.01
<i>Sus scrofa</i>	Serine protease inhibitor-2	X16362	Exon 2	-2.88 ± 0.47	Exons 3, 4, 5	-0.84 ± 0.74	-2.04	0.02
	Cytochrome P450 steroid 21-hydroxylase	M83939	Exon 10	0.63 ± 0.68	Exons 1–9	-0.60 ± 0.52	1.23	0.145
<i>Homo sapiens</i>	Tenascin-X	AF019413	Exon 45	3.25 ± 1.00	Exons 34–44	0.76 ± 0.47	2.50	0.042

^a Base order-dependent stem-loop (folding) potentials (FORS-D values), were derived by subtracting base composition-dependent stem-loop potentials (FORS-M values), from total stem-loop potentials (FONS values).

^b T4 phage ORF 30.3 overlaps ORFs 30.4, 30.3' and 30.2. ORF Y overlaps ORF 30.2. G4 phage ORFs A and D overlap ORFs B, K, and E. *Achlya klebsiana* glutamate dehydrogenase exon 10 overlaps heat shock protein 70 (U02504). *Drosophila melanogaster* fructose 1,6-bisphosphate aldolase exon 4A overlaps exon 4B. *Rattus norvegicus* gonadotrophic-releasing hormone exons 2 and 3 overlap exons 0, 1 and 4 of the SH-4 protein gene. Exon 2 of serine proteinase inhibitor-2 overlaps exon 1 of a mitochondrial cytochrome P-450 C27-steroid hydroxylase (U17375). Exon 10 of *Sus scrofa* cytochrome P-450 C21-steroid hydroxylase overlaps exon 39 of the putative swine tenascin-X gene. Exon 45 of the human tenascin-X gene overlaps exon 10 of the adrenal C21-steroid hydroxylase gene (u24488).

^c Average of eight values in this column is 1.60 ± 0.58 (*P* < 0.03). The last two values involve the same genes in different species. If *Sus scrofa* is removed from the analysis, the average is 1.65 ± 0.66 (*P* < 0.05).

^d Probability that not significantly different from zero (unpaired *t*-test).

windows is computationally intensive, we sampled windows at 25 base intervals, and were usually able to collect sufficient samples to ensure P values of the order of, or less than, 0.05.

2.4. Overlapping windows

For some purposes it may be necessary to avoid overlapping sequence windows ‘to ensure the independence of sampling points’ (Alvarez-Valin et al., 2000). By virtue of collecting our samples in a common region (exon or ORF) it follows that each window overlaps several others, but this is not relevant to the question examined; namely, what is the probability that a sampling of windows from an overlap segment differs from a similar sampling of windows from a non-overlap segment?

It may not be intuitively obvious that windows which overlap other windows can be treated independently for statistical purposes. Consider the problem of whether an area between an interval on the X -axis and the curve connecting one set of experimental points is significantly different from an area between an equal interval on the X -axis and the curve connecting another set of experimental points. To make this distinction, we can evaluate each area by summing the areas of rectangles each with a height (Y -axis) corresponding to the data value and with a base corresponding to the interval between data points on the X -axis. Since we fix this interval, we can just take one set of Y values and compare this with the other set of Y values.

As example, imagine that a car drives from A to B in a certain time and comes to a halt. Then in the same time it drives from B to C. The car accelerates from A, and then decelerates to B, then accelerates from B and decelerates to C. We have velocity values of uncertain accuracy for different time points. We want to know if the distance AB is different from the distance BC. Each distance equals the area under the corresponding plot of velocity (Y -axis) against time (X -axis). If velocity readings are taken at regular intervals, we can take the average of the velocity readings for the distance AB and compare with the average of the velocity readings for the distance BC. If the two averages are significantly different, then the two distances are significantly different. However, the velocity readings are not independent. Having already accelerated to a velocity of 20 km/h in one time interval, the car is set to accelerate to a velocity of 40 km/h in the next time interval. This interdependence is not relevant to the question whether the sample of velocity values for the distance AB, is significantly different from the sample of velocity values for the distance BC.

3. Results

A region required simultaneously to encode two proteins either in different reading frames in one strand (unidirectional transcription), or in complementary strands (bi-direc-

tional transcription), should be less able to accommodate stem-loop potential.

A member of the tumor necrosis factor receptor family, the human B cell maturation protein encoded by the *BCMA* gene, has an antisense transcript encoding a putative 115 amino acid protein. Fig. 1 shows that the overall folding potential (FONS) is greater (i.e. high negative values) in introns and flanking regions, than in exons. The potential is contributed partly by base composition (FORS-M), and partly by base order (FORS-D). As in previous work (Forsdyke, 1995a,b), the difference between introns and exons is far from dramatic. There appears to be much ‘noise’, and the case is best made *statistically*, not *visually*. The average base order-dependent stem-loop potential of the two introns (average FORS-D of -2.94 ± 0.56), is significantly greater ($P < 0.004$) than that of the three exons (average FORS-D of -0.24 ± 0.56). However, this does not establish that overlapping coding functions have influenced the difference. For this, it is necessary to examine sequences in GenBank that contain both overlapping and non-overlapping exons in proximity.

The human HLA class III region contains a gene encoding the extracellular matrix protein tenascin-X, the antisense strand of which contains the gene encoding adrenal steroid 21-hydroxylase (*P450-C21*). Fig. 2 shows that, where exon 45 of the former overlaps exon 10 of the latter, FORS-D values (measuring base order-dependent stem-loop potential) are positive. The figure also illustrates some of the difficulties in this type of analysis. In the region of exons 40–45 there is, as expected, an alternation between low and high base order-dependent stem-loop potentials, with low potentials corresponding with exons, and high potentials corresponding with introns. However, this is not so apparent in the case of exons 34–39. Indeed the largest exon (34) has high base order-dependent stem-loop potential, with a region of low potential (positive FORS-D values) in its 3′ flank. This indicates that the segment of the protein corresponding to exon 34 can be encoded without impairing stem-loop potential, whereas in the 3′ flank there might be a conserved region, perhaps playing a regulatory role, whose sequence cannot accommodate stem-loop potential (Forsdyke, 1995a).

To demonstrate that exon 45 was under significantly greater constraint than most neighboring exons, the average FORS-D values of windows whose centers overlapped the exon (3.25 kcal/mol) was compared with the average FORS-D values of windows whose centers overlapped exons 34–44 (0.76 kcal/mol). The difference was 2.5 kcal/mol ($P = 0.042$). This approach was applied to seven other genes from a variety of organisms. In all cases except one, differences were positive (penultimate column of Table 1). The average difference was 1.60 ± 0.58 ($P < 0.03$).

The exception is shown in Fig. 3. In this case, the overlap is between a large exon (exon 2) of the gene encoding rat serine proteinase inhibitor-2, and exon 1 of a steroid 27-hydroxylase. It is possible that exon 2 is large (i.e. has not been split in the course of evolution) because it has suffi-

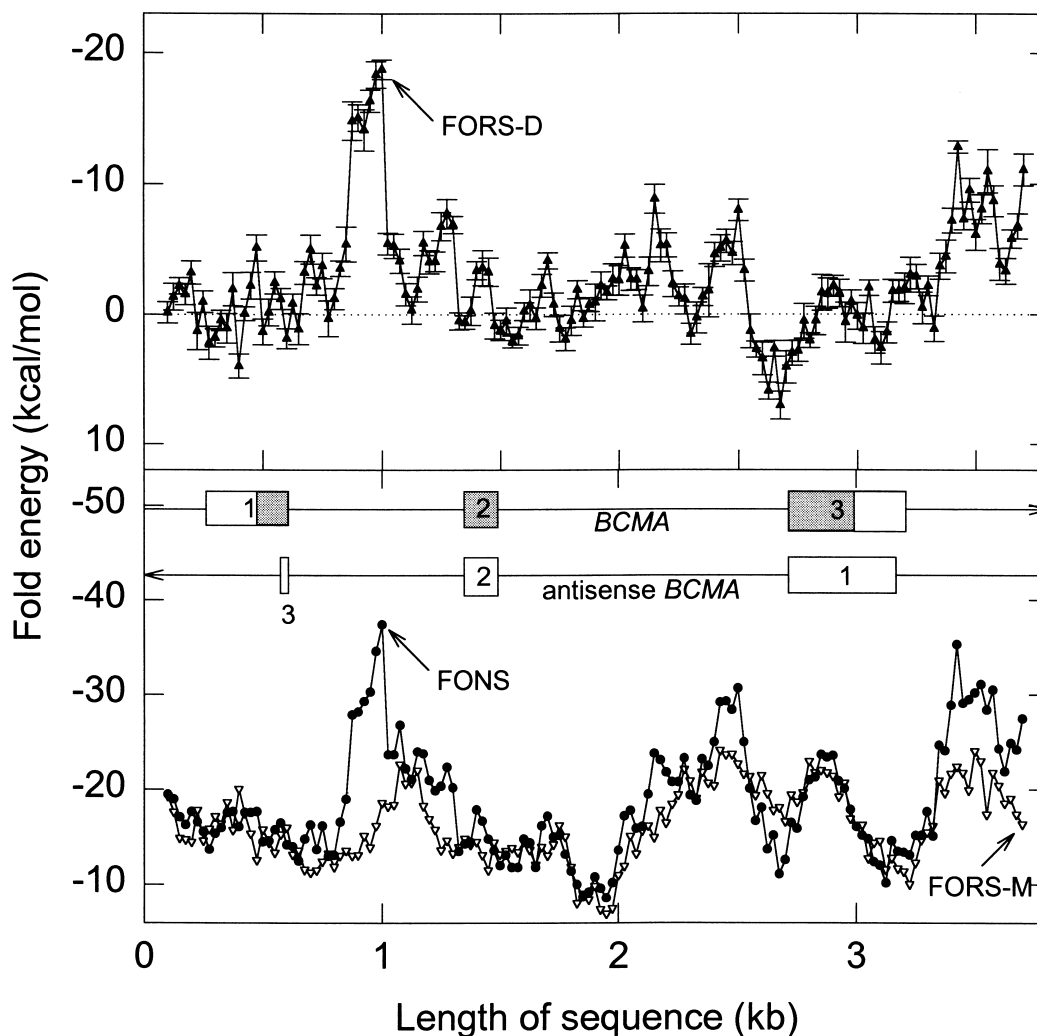


Fig. 1. Folding energy minimization values (FONS, FORS-M; lower) and differences (FORS-D; upper) for the top strand of the 3802 nt gene (*BCMA*) encoding human B cell maturation protein (GenBank accession number Z29574). The overlapping gene in the bottom strand encodes a putative 115 amino acid protein. Exons are shown as numbered boxes with regions encoding the *BCMA* protein in grey. Horizontal arrows indicate transcription directions. Each data point corresponds to the middle of a 200 nt sequence window.

cient flexibility in choice of amino acids and codons to accommodate both to stem-loop potential and to the coding needs of the antisense strand.

4. Discussion

4.1. Function of introns?

In the 1950s it was found that much RNA freshly synthesized in the nucleus was locally degraded and did not reach the cytoplasm. In the 1960s eukaryotic rRNAs were found to be synthesized as long precursor RNAs ('heterogenous nuclear RNAs') which were subsequently processed by the intranuclear removal and degradation of apparently functionless internal 'spacer' sequences (Harris, 1994). When prokaryotic rRNAs were found to be more compactly organized, it was appropriate to ask whether the first rRNAs to

evolve had the spacer sequences, which subsequently decreased in prokaryotes, or whether the spacer sequences were later acquired in eukaryotes.

A similar processing was later found to apply to eukaryotic mRNAs. After transcription internal sequences ('introns') were removed, and what remained in the processed mRNA constituted the 'exons'. Since the phenomenon had first been noted in the case of rRNA genes, which were not translated into a protein product, it was not surprising that introns were found in other RNAs with no protein product (Pfeifer and Tilghman, 1994), as well as in non-coding parts of mRNAs which had a protein product. Thus, introns interrupted genetic information *per se*, not just protein-encoding information, and it was difficult to associate exons with domains of protein structure or function (Weber and Kabsch, 1994; Stoltzfus et al., 1994). The same questions remained. Were introns 'early' or 'late'? What function(s), if any, did introns have?

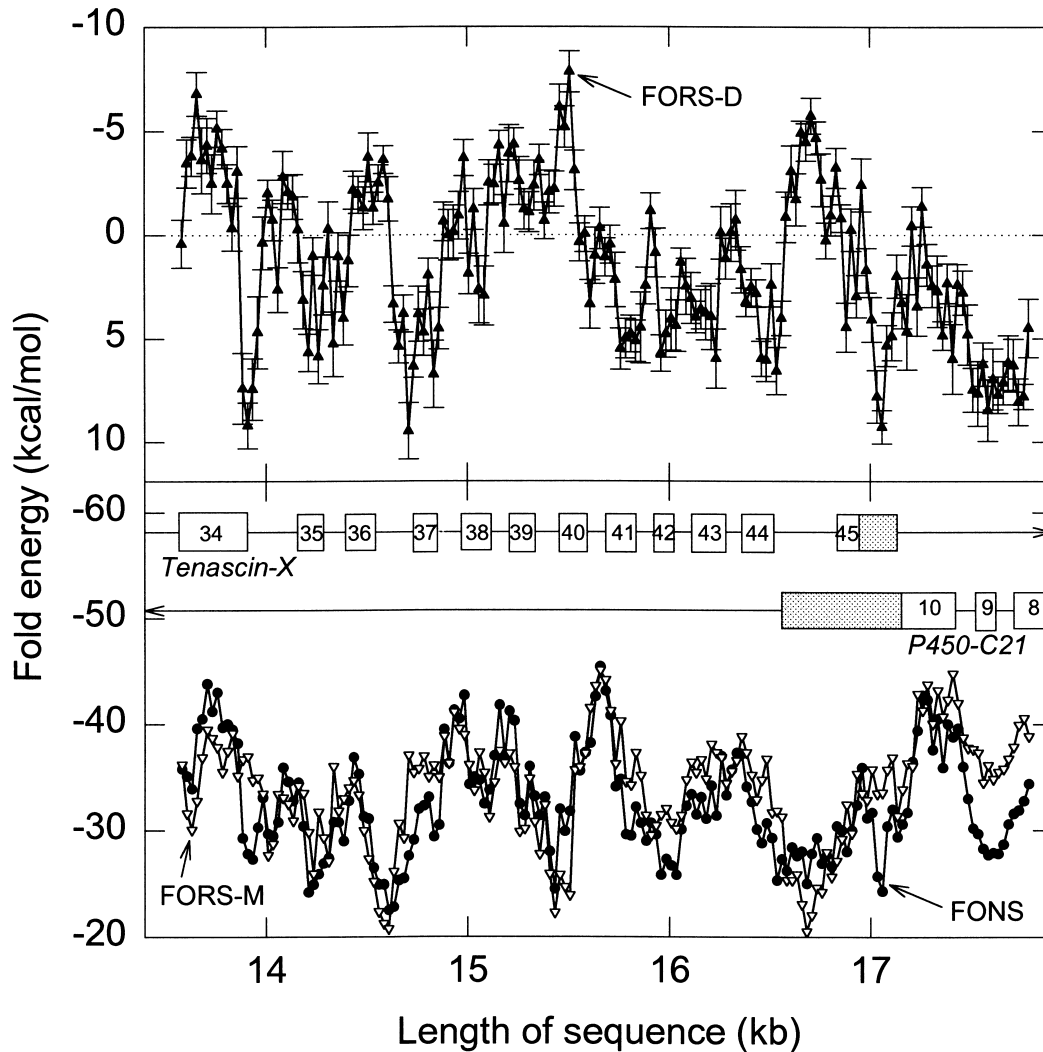


Fig. 2. Folding energy minimization values for a 4400 nt 3' segment of the human gene encoding tenascin-X (Accession number AF019413), which overlaps the 3' end of the gene *P450-C21* encoding cytochrome P450 21-hydroxylase. Grey shaded areas indicate 3' untranslated regions of terminal exons. For further details see the legend to Fig. 1.

If bacteria could exist without introns, then perhaps introns had no function. Alternatively, whatever function introns had, either was not necessary in bacteria, or might be achieved in other ways. Since members of many bacterial species appeared to be under intense pressure to streamline their genomes to facilitate rapid replication, if it were possible they would have dispensed with any preexisting introns and/or would have been reluctant to acquire them. On the other hand, if introns had a function and/or did not present too great a selective burden, eukaryotes would have tended to retain preexisting introns, or could have acquired them. Knowing the function of introns seemed critical for sorting out these issues.

4.2. Error-correction

Some principles to guide investigation of a possible error-checking role for introns were presented (Forsdyke, 1981).

Although the mechanism may be different to that originally proposed (Liebovitch et al., 1996), the present work is part of a growing body of evidence that introns play such a role (Forsdyke, 1995a–d, 1996a,b, 1998; Heximer et al., 1996). It appears possible that the order of bases in nucleic acids have been under evolutionary pressure to develop the potential to form stem-loop structures which would facilitate 'in-series' or 'in-parallel' error-correction by recombination (Forsdyke, 1996b; Bell and Forsdyke, 1999; Forsdyke, 2001a,b).

4.3. Evolution rate determines positive or negative correlation

Whereas base composition tends to be a characteristic of entire genomes or large genome sectors, base order tends to be a local characteristic, like the ability to encode a protein. Despite the background 'noise', which makes statistical

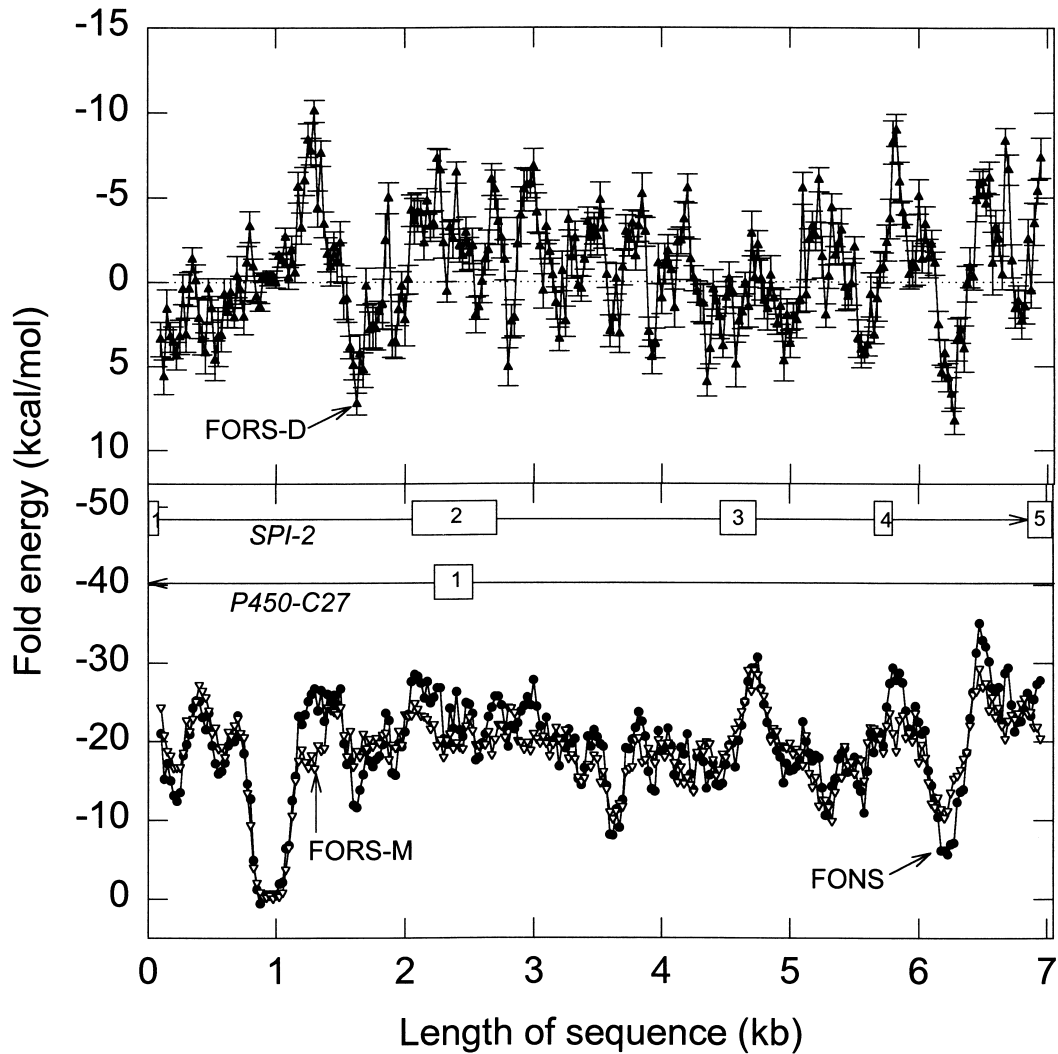


Fig. 3. Folding energy minimization values for a 7050 nt segment of the rat *SPI-2* gene encoding serine proteinase inhibitor-2 (Accession number X16362), which overlaps the 5' end of a gene encoding a steroid 27-hydroxylase. For further details see the legend to Fig. 1.

analysis essential, base order-dependent stem-loop potential can be a sensitive indicator of functional conflict acting at the local level between a potential error-correcting function and a protein-encoding function. One example of such conflict is that of genes evolving rapidly under *positive* Darwinian selection. Here there is a *negative* correlation (reciprocal relationship) between base order-dependent stem-loop potential and sequence variability (Forsdyke, 1995b,d, 1996a). Sequences varying *rapidly* in response to powerful environmental selective forces ('arms races') appear unable simultaneously to encode base sequences favoring the elaboration of a higher order structure of a type which might mediate meiotic chromosomal interactions. The latter function must either be relegated to those less rapidly evolving protein-encoding sequences (or parts of such sequences) where there is some flexibility in codon or amino acid choice (the main option in compact genomes), or to non-protein-encoding regions (introns and non-genic DNA).

On the other hand, in the case of *slowly* evolving sequences, there may be a *direct* relationship between base order-dependent stem-loop potential and sequence variability. The demand of faithful reproduction of a protein, with *negative* Darwinian selection of individuals with mutations affecting the functionally most important parts of the protein, leaves the co-encoding of stem-loops to regions encoding functionally less important parts of the protein (such as the protein surface; Alvarez-Valin et al., 2000), and to non-protein-encoding regions (introns, non-genic DNA). High stem-loop potential can then correlate *positively* with high substitution rates (variability) when homologous sequences from different species are compared (Heximer et al., 1996). In the case of compact genomes with no intron option, the need to form stem-loops could sometimes have over-ridden the demands of protein structure and function, so that less-than-perfect proteins resulted (Ball, 1973). In this case the negative selection of individuals because of impaired stem-loop forming ability, would

have been greater than the negative selection resulting from impaired protein function.

4.4. *Demonstration of conflict*

High base order-dependent stem-loop potential (assessed here as high negative FORS-D values), appears as the ‘default’ state of genomes, which probably arose in the early RNA world (Forsdyke, 1995a,c; Heximer et al., 1996). We have sought here ‘echoes’ of the evolutionary ‘big bang’ that occurred when protein-encoding capacity arose to compete with stem-loop potential for genome space. Despite the ‘noise’ associated with DNA’s evolutionary role as a ‘channel’ for multiple forms of information, if the need to encode a protein were to conflict with the need to encode stem-loops, then it would be predicted that stem-loop potential would detectably decrease (i.e. FORS-D values would tend to become positive) in modern protein-encoding regions (e.g. exons). Both the degeneracy of the genetic code allowing changes in codons without changing the corresponding amino acid, and exchanges between amino acids of similar function, might mask the conflict. If so, an exon could be large. If not so, a potentially large exon would be split into smaller exons by the interspersions of introns with sequences of high stem-loop potential.

In initial studies, a decrease in base order-dependent stem-loop potential (FORS-D) in exons was demonstrated, but sometimes required removing the 3′ terminal exon (often the largest) from the analysis (Forsdyke, 1995a). It was found that the relationship was best shown by genes which were evolving rapidly under positive Darwinian selection. These included the snake venom phospholipases (Forsdyke, 1995b), retroviral genomes (Forsdyke, 1995d), MHC proteins (Forsdyke, 1996a), and troponin C (Forsdyke, 1996b). The extreme pressure to adapt protein function appeared to countermand accommodation of base order-dependent stem-loop potential.

In the present work we examined the effect of the extreme demand made on the function when genes overlap. Although not visually obvious, from statistical analyses we found that base order-dependent stem-loop potential was usually countermanded (Figs. 1 and 2, Table 1), with one unexplained exception (Fig. 3). Collected from a wide variety of species, we believe that the genes studied are likely to be generally representative of overlapping genes. Thus, consistent with the conflict hypothesis under investigation, our results support the generalization that regions of gene overlap support stem-loop potential less than regions of non-overlap.

4.5. *Gene size and role of junk DNA*

Regions where the constraints of classical Darwinian negative and positive selection, or of the encoding of overlapping genes, would not permit accommodation of DNA stem-loop potential, would be expected to be rich in introns. The more severe the constraints, the more introns there

would be, and the longer would be the length of DNA occupied by the gene (Naora and Deacon, 1982). The adaptive value of recombination involving stem-loops could have been manifest in the early RNA world, so that protein-encoding capacity would have intruded on genomes already adapted for stem-loop formation. In this respect, introns would have been ‘early’, not ‘late’. With each speciation event there should be changes in stem-loop potential (Forsdyke, 1996b, 1998), with the possibility of further changes in intron number and distribution. However, although this provides a possible explanation for the origin of introns, it does not appear to explain why introns are sometimes of extremely large size. A full understanding of factors affecting intron size may require a better understanding of the functions of intronic and non-genic DNA other than the stem-loop function (Forsdyke and Mortimer, 2000).

4.6. *Purine-loading should promote intron formation*

Another problem relates to the existence of introns in genes with no protein product. Instead the products are RNAs, which have specific functions dependent on their secondary structures (usually selected for at the cytoplasmic level). In broad features we find that computer-derived secondary structures for an RNA molecule (using pairing energy tables for RNA bases), are similar to the structures derived for the corresponding DNA (using pairing energy tables for DNA bases; D. R. Forsdyke, unpublished work). If such stem-loop structures were sufficient for function at the DNA level, then there would be no need for introns in genes for non-protein-encoding RNAs. Thus we surmise that stem-loop structures which suffice for function at the RNA level, do not suffice for function at the DNA level. Since patterns of RNA stem-loops are influenced by the purine-loading of loops (the selective force for which probably operates at the mRNA level; Forsdyke and Mortimer, 2000), then purine-loading pressure (which would constrain DNA-specific stem-loop patterns in exons) should support stem-loop pressure in provoking the splitting of large exons.

Acknowledgements

We thank J. Gerlach for assistance with computer configuration. Academic Press and Elsevier Science gave permissions for the inclusion of full-text versions of relevant preceding papers at our internet site (<http://post.queensu.ca/~forsdyke/bioinfor.htm>).

References

- Alvarez-Valin, F., Tort, J.F., Bernardi, G., 2000. Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from *Leishmania*. *Genetics* 155, 1683–1692.
- Ball, L.A., 1973. Mutual influence of the secondary structure and information content of a messenger RNA. *J. Theor. Biol.* 41, 243–247.

- Bell, S.J., Forsdyke, D.R., 1999. Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. Theor. Biol.* 197, 63–76.
- Bronson, E.C., Anderson, J.N., 1994. Nucleotide composition as a driving force in the evolution of retroviruses. *J. Mol. Evol.* 38, 506–532.
- Carlson, E.A., 1981. *Genes, Radiation and Society, The Life and Work of H.J. Muller*. Cornell University Press, Ithaca, pp. 390–392.
- Crick, F., 1971. General model for the chromosomes of higher organisms. *Nature* 234, 25–27.
- Doyle, G.G., 1978. A general theory of chromosome pairing based on the palindromic model of Sobell with modifications and amplification. *J. Theor. Biol.* 70, 171–184.
- Eguchi, Y., Itoh, T., Tomizawa, J., 1991. Antisense RNA. *Annu. Rev. Biochem.* 60, 631–652.
- Forsdyke, D.R., 1981. Are Introns In-series Error-detecting Sequences? *J. Theor. Biol.* 93, 861–866.
- Forsdyke, D.R., 1995a. A stem-loop 'kissing' model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.* 12, 949–958.
- Forsdyke, D.R., 1995b. Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes. An application of FORS-D analysis. *Mol. Biol. Evol.* 12, 1157–1165.
- Forsdyke, D.R., 1995c. Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *J. Mol. Evol.* 41, 573–581.
- Forsdyke, D.R., 1995d. Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *J. Mol. Evol.* 41, 1022–1037.
- Forsdyke, D.R., 1996a. Stem-loop potential in MHC genes: a new way of evaluating positive Darwinian selection. *Immunogenetics* 43, 182–189.
- Forsdyke, D.R., 1996b. Different biological species 'broadcast' their DNAs at different (C + G)% 'wavelengths'. *J. Theor. Biol.* 178, 405–417.
- Forsdyke, D.R., 1998. An alternative way of thinking about stem-loops in DNA. A case study of the human *GOS2* gene. *J. Theor. Biol.* 192, 489–504.
- Forsdyke, D.R., 2001a. Functional constraint and molecular evolution. *Encyclopedia of Life Sciences*. Macmillan Reference Ltd, London.
- Forsdyke, D.R., 2001b. *The Origin of Species, Revisited*. McGill-Queen's University Press, Montreal.
- Forsdyke, D.R., Mortimer, J.R., 2000. Chargaff's legacy. *Gene* 261, 127–137.
- Harris, H., 1994. An RNA heresy in the fifties. *Trends. Biochem. Sci.* 19, 303–305.
- Heximer, S.P., Cristillo, A.D., Russell, L., Forsdyke, D.R., 1996. Sequence analysis and expression in cultured lymphocytes of the human *FOSB* gene (*GOS3*). *DNA Cell Biol.* 15, 1025–1038.
- Le, S.H., Maizel, J.V., 1989. A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.* 138, 495–510.
- Liebovitch, L.S., Tao, Y., Todorov, A.T., Levine, L., 1996. Is there an error-correcting code in the base sequence of DNA? *Biophys. J.* 71, 1539–1544.
- Muller, H.J., 1922. Variation due to change in the individual gene. *Am. Nat.* 56, 32–50.
- Naora, H., Deacon, N.J., 1982. Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. *Proc. Natl. Acad. Sci. USA* 79, 6196–6200.
- Pfeifer, K., Tilghman, S.M., 1994. Allele-specific gene expression in mammals: the curious case of imprinted RNAs. *Genes Dev.* 8, 1867–1874.
- Salser, W., 1970. Discussion. *Cold Spring Harbor Symposium. Quant. Biol.* 35, 19.
- Santalucia, J., Allawi, H.T., Seneviratne, P.A., 1996. Improved nearest neighbour parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555–3562.
- Seffens, W., Digby, D., 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* 27, 1578–1584.
- Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M., Doolittle, W.F., 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* 265, 202–207.
- Weber, K., Kabsch, W., 1994. Intron positions in actin genes seem unrelated to the secondary structure of protein. *EMBO J.* 13, 1280–1286.
- Winge, Ö., 1917. The chromosomes, their number and general importance. *Compte. Rend. Trav. Lab. Carlsberg* 13, 131–275.