A Stem-Loop “Kissing” Model for the Initiation of Recombination and the Origin of Introns

D. R. Forsdyke
Department of Biochemistry, Queen's University

Mutations which improve the efficiency of recombination should affect either the proteins which mediate recombination or their substrate, DNA itself. The former mutations would be localized to a few sites. The latter would be dispersed. Studies of hybridization between RNA molecules have suggested that recombination may be initiated by a homology search involving the “kissing” of the tips of stem loops. This predicts that, in the absence of other constraints, mutations which assist the formation of stem loops would be favored. From comparisons of the folding of normal and shuffled DNA sequences, I present evidence for an evolutionary selection pressure to distribute stem loops generally throughout genomes. I propose that this early pressure came into conflict with later local pressures to impose information concerning specific function. The conflict was accommodated by permitting sections of DNA concerned with a specific function to evolve in dispersed segments. Traces of the conflict seem to be present in some modern intron-containing genes. Thus, introns may have allowed the interspersing of selectively advantageous stem loops in coding regions of DNA.

Introduction

The possession of efficient mechanisms for homologous recombination would appear to be evolutionarily advantageous (Bernstein and Bernstein 1991). Thus, mutations in DNA which improve the efficiency either of the proteins concerned with recombination, or of the DNA to act as a target for these proteins, would be evolutionarily advantageous. Mutations improving the efficiency of proteins would be expected to be localized to the region of the corresponding genes. However, mutations improving the ability of DNA to act as a recombination substrate would be expected to be widely distributed.

The view that prior formation of a synaptonemal complex is essential for the initiation of meiotic recombination has recently lost ground to the view that an initial sequence-based homology search precedes chromosome pairing (Hawley and Arbel 1993). In 1971 Crick advanced the “unpairing postulate” that supercoiled duplex DNA in chromatin would form double-stranded stem-loop structures. At the tips of these structures individual strands would unpair, providing the opportunity for interactions with single-stranded DNA at the tips of similar stem loops on other chromosomes. This would allow a homology search and the pairing of homologous chromosomes during meiosis. There was no clear prediction that the primary sequence could have evolved to facilitate this.

Subsequent studies of the mechanism of hybridization between the sense and antisense RNAs involved in regulation of the replication of plasmid ColE1 indicated that an initial homology search between the complementary single-stranded RNAs involves weak, reversible, transient “kissing” interactions between the tips of stem-loop structures (Tomizawa 1984). Supporting evidence for a role of stem loops in recombination between RNA genomes was provided by Romanova et al. (1986). This raised the possibility that similar stem-loop structures, generated when duplex DNA adopts a cruciform configuration, are involved in the initial homology search during meiosis (Sobell 1972; Wagner and Radman 1975; Kleckner and Weiner 1993).

If stem-loop formation in duplex DNA facilitates the initiation of homologous recombination, then there should have been a mutational pressure to produce dispersed sets of complementary oligonucleotide pairs. It is known that the frequencies of members of complementary oligonucleotide pairs are similar in DNA (Pradhu 1993; Forsdyke 1995b). However, it is not known whether the complementary pairs are sufficiently colocalized to permit extensive stem-loop formation.
An algorithm for the computation of "statistically significant" stem-loop potential was introduced by Le and Maizel (1989). This involved comparison of the folding into stem-loop structures of small windows in nucleic acid sequences with the folding of the same windows after randomizing base order. In this paper, their approach is adapted for the determination of the distribution of stem loops in long DNA molecules. I first examine the distribution of stem-loop potential in some of the long DNA sequences (e.g., 68 kb) which have recently become available. The results support the single-stranded DNA stem-loop "kissing" model. I then speculate that an early genome-wide evolutionary pressure for the formation of dispersed stem-loop structures came into conflict with later local pressures to encode specific functions affecting phenotype. Finally, I present a survey of a variety of modern genes to see if traces of the conflict are still discernible. The results cast a new light on both the exon/intron problem (Gilbert and Glynias 1993; Stoltzfus et al. 1994) and the problem of why different genomes or genome compartments have distinct (G+C)/(A+T) ratios (Wyatt 1952; Bernardi 1989). The latter is discussed more fully elsewhere (unpublished data).

Methods

The secondary structure of DNA in single-stranded form is very sensitive to small changes in sequence (Orita et al. 1989). Provided the length is not excessive, any such sequence can be analyzed using computer programs such as FOLD (Zuker 1989) to arrive at a theoretical optimum secondary structure of minimum free energy. Because of the greater strength of GC bonds relative to AT bonds, a GC-rich sequence tends to have a more stable structure than an AT-rich sequence of the same length. However, it does not follow that the most stable structures are likely to be of most local functional relevance. The bases in a segment might show poor complementarity, but if the few complementary pairs were GC pairs, the stability of the folded molecule might be quite high. Base composition is a genome, or genome sector (isochore), "strategy," which has a major influence on codon choice, particularly of synonymous codons (Grantham et al. 1980). Base composition is not a local "strategy." Codons are a local "strategy." Confidence that a given secondary structure is of local functional relevance is greater if it can be shown that the sequence has accepted mutations which enhance stem-loop formation (i.e., that the actual sequence of bases, in addition to base composition, has contributed to secondary structure stability). In principle, this should be possible by comparing the folding of a natural sequence with that of randomized versions of the same sequence. A natural sequence is but one member of a large set of possible sequences with the same base composition. The average characteristics of this set can be arrived at by randomizing the natural sequence. Randomization (shuffling) destroys information present in the primary sequence (base order), without changing base composition or sequence length. Thus, provided length is kept constant, average characteristics reflect base composition alone.

Programs of the Genetics Computer Group Inc. (Gribskov and Devereux 1991) were made available online through the services of the Molecular Biology Data Service of the National Research Council, Ottawa, which includes access to a Silicon Graphics Challenge XL computer. Sequences were randomized using the program SHUFFLE. The outputs from the latter could be used directly by the program FOLD (Zuker 1989), which finds a secondary structure of minimum free energy using the energy values for base stacking and loop destabilization assigned by Turner et al. (1988). This program was designed for the study of secondary structure in nonsupercoiled, single-stranded RNA molecules. It is assumed here that it is applicable to DNA. This assumption is not entirely valid. For example, the 2' hydroxyl group in RNA can contribute of the order of 1 kcal/mol to helix stability (Turner and Bvilaqua 1993). Since the present study is mainly concerned with differences in fold energies rather than absolute values, differences between RNA and DNA are considered of minor importance. A unix script program, SHUFFOLD, was written to determine the minimum free energy of folding of natural and shuffled versions of successive overlapping 200-nt windows from nucleic acid sequences. Another program, STATS, was written to subject the output from SHUFFOLD to statistical analysis in the Minitab system (Ryan and Joiner 1994).

For each 200-nt window, FOLD first determines the minimum free energy value for folding of the natural sequence (FONS value). This is a function of both base composition and base order and measures the "total stem-loop potential" of a region. Then 10 random sequences are generated from the same window, and each randomized sequence is submitted to FOLD. The mean minimum free energy value for the 10 sequences (FORS-M value) provides a measure of the contribution of base composition alone to the stem-loop potential ("base composition-determined stem-loop potential"). Since the FONS value is usually more negative than the FORS-M value, the difference between the two values (FORS-M less FONS) is usually positive. This difference (the "FORS-D" value) provides a measure of the contribution of base order alone to the stem-loop potential. Thus, a positive FORS-D value defines and quantitates the "base order-determined stem-loop potential." This closely corresponds to the "segment score" of Le and
negative FORS-D values are widely dispersed in segments from small, medium, and large human chromosomes. Long DNA fragments from Drosophila melanogaster, Escherichia coli, and bacteriophage lambda also showed significant positive average FORS-D values (Forsdyke 1995b).

The dispersal of positive FORS-D values throughout genomes indicates a general evolutionary pressure on DNA molecules to accept mutations which would promote stem-loop formation by affecting base order. This supports the stem-loop "kissing" model for recombination as outlined in the Introduction. FORS-D profiles show negative values in some regions, suggesting conflicts between a general pressure on a gene to optimize its folding propensity and local pressures for some function.

Potential conflicts between general and local pressures on evolving DNA molecules are summarized in figure 2. The two rows of downward-pointing arrows symbolize pressures acting throughout the genome. One of these is the general pressure to set the \((G+C)/(A+T)\) ratio (base composition) at a particular level in a particular genome or genome segment ("GC/AT pressure"); see Discussion. The second is the pressure, quantified as the FORS-D, to generate sequences with complementary secondary structure potential.

Maizel (1989), which is used to assess "statistically significant" stem-loop potential. A negative FORS-D value in a region may mean that base order has been adapted to serve some other potential.

Results

Positive FORS-D Values Are Dispersed in Genomic DNA

A study of folding energies was carried out on one strand of the 68-kb human chromosome 19 segment HUMMMDDBC (GenBank name). If there had been a genome-wide pressure to maximize the potential to form secondary structures, then positive FORS-D values would be expected. To get a general impression of the folding propensity, the FOLD program was applied to isolated 200-nt windows, at 1-kb intervals. The FONS and FORS-M profiles tend to follow each other (fig. 1B). Positive FORS-D values are widely dispersed and greatly exceed negative FORS-D values (fig. 1A; average 4.37 ± 0.90 kcal/mol). Similar results are obtained for the HUMHBB segment from chromosome 11 (average 4.49 ± 1.34 kcal/mol for 15 windows at 5-kb intervals) and for the HUMHDABC segment from chromosome 4 (average 3.59 ± 1.62 kcal/mol for 12 windows at 5-kb intervals). Thus positive FORS-D values are widely dispersed in segments from small, medium, and large human chromosomes.
FIG. 3.—Fold energy minimization values (FORS-M, FONS) and differences (FORS-D) for A, the 4,102-nt sequence containing the GOS19-1 cytokine gene (GenBank name HUMGOS19A), and B, the corresponding cDNA. Each data point corresponds to the middle of a 200-nt window. Each window overlaps the preceding window by 150 nt, except that the first of the 41 overlapping windows of the genomic sequence spans nt 1-194. Thus windows, in order, correspond to nt 1-194, nt 45-244, nt 95-294, and so forth. This determines that the beginning of exon 1 corresponds to a new window (nt 1,995-2,194). In B there are 14 overlapping windows which begin with nt 1. The last window of the 778-nt cDNA sequence spans nt 601-778. The three exons are shown as open boxes. Vertical dashed lines in A indicate, from left to right, the beginning of exon 1, the beginning of the protein-encoding region, the end of the protein-coding region, and the end of exon 3. Vertical dashed lines in B show where introns have been removed.

mentary oligonucleotides situated so as to favor the formation of stem loops. These two sets of arrows are pointing in the same direction, indicating that there is little conflict between them. The upward-pointing arrows are in distinct regions, symbolizing localized evolutionary pressure for the encoding of specific function. Here there is a conflict. A sequence required to encode a protein might not simultaneously be able to optimize its folding propensity.

FORS-D Values of Intron-Containing Genes

To seek traces of this postulated evolutionary conflict between a genome-wide FORS-D pressure and local pressures related to individual gene function, the folding patterns of some modern genes were examined. Fold energy plots were prepared using consecutive 200-nt windows with 150-nt overlaps. The human cytokine-encoding gene GOS19-1 is one of a series of potential lymphocyte G0/G1 switch regulatory genes ("GOS genes") which have been sequenced in my laboratory (Blum et al. 1990). The gene has three exons distributed over a region of 1,886 nt, and the cDNA derived by splicing is 778 nt in length. As noted above (fig. 1), FONS and FORS-M profiles tend to resemble each other (fig. 3A, lower panel). FORS-D values are generally positive in the gene; however, distinct local decreases occur in certain regions (fig. 3A, upper panel). When the introns are removed to generate the cDNA, decreased FORS-D values tend to be concentrated together, and general positivity is less evident (fig. 3B, upper panel). An extremely low FORS-D value in the immediate 5' flank of the gene corresponds to a region rich in potential regulatory motifs, which is highly conserved between GOS19-1 and its murine homolog (Russell and Forsdyke 1993). This supports the idea of a conflict between a general evolutionary pressure for stem-loop potential and local pressure for function. The very high positive FORS-D values (>20 kcal/mol) in the 5' flank of the GOS19-1 gene correspond to a potentially "foreign" AT-rich minisatellite-like element containing four 22-nt tandem repeats and two inverted repeats. (Only 1 of the 22-nt repeats is present in the GOS19-2 gene; Blum et al. 1990.) Low FORS-D values in the regions of the second and third exons are also consistent with the conflict hypothesis; however, values are high in the signal peptide-encoding first exon (fig. 3A).

Another GOS gene (GOS7) corresponds to the oncogene c-fos, which has four exons (fig. 4). Each exon is associated with a region of negative FORS-D value, whereas the three introns tend to have high FORS-D values. The negative values of the exons are not reflected in the cDNA, which shows no consistent tendency to low values (fig. 4B). The coding part of the fourth exon is long (642 nt). Although the 5' part of the exon is associated with a region of low FORS-D value, the rest of
the coding region is associated with high FORS-D values. Thus, here there appears to be no conflict between pressures for the evolution of protein-encoding function and for the potential to form stem loops. The 3' part of the 3' noncoding region has relatively low FORS-D values, consistent with the known functional role of this region in controlling mRNA stability.

FORS-D Values Correlate Negatively with Exon Overlap

To obtain an objective measure of the tendency of small exons to associate with low FORS-D values, each of the 200-nt windows for which FORS-D values were determined was scored for its percentage of exon overlap. Figure 5A shows a linear regression plot for the three exons of the G0S19-1 gene. Consecutive windows extended from the first 5' window overlapping the first exon to the last 3' window overlapping the last exon. Although there was much scatter of data points, the slope of the least-squares regression line was significantly greater than zero \( (P = 0.048 \text{ that the slope is not greater than zero}) \). Figure 5B and C show similar data for the oncogenes c-fos and p53, except that the long last exons were omitted. Both slopes were significantly greater than zero \( (P = 0.018, \text{c-fos}; P = 0.002, \text{p53}) \). The corresponding \( P \) values when the last exons were included were 0.752 and 0.048. (A justification for omitting last exons is given later.)

The above analyses were extended to a broader spectrum of genes, and the results are shown in table 1. Five of 12 human genes had significant \( P \) values \(< 0.05 \text{ that the slope is not significantly greater than zero} \). The one mouse gene studied had a significant \( P \) value. The one plant gene studied (TOMBIPGRP) did not have a significant \( P \) value. The most significant \( P \) value was obtained with a gene (encoding troponin C), believed to have been under positive Darwinian selection (Ohta 1994). Table 2 compares average FORS-D values of gene-containing segments (table 1) with those of the corresponding cDNAs. In most cases average FORS-D values for the genomic segments were greater \( (+0.71 \pm 0.33; P < 0.05 \text{ by Student’s } t\text{-test}) \).

Discussion

The Competing Needs of the Early Replicators

The winner in the competition between replicator molecules in early evolution would have been the replicator which could most efficiently balance the needs for (1) rapid replication, (2) accurate replication, and (3) stability. Intimately linked with the latter two needs would be the need to carry out repair of malreplicated or damaged molecules. In an early “RNA world,” the ability to shuffle damaged segments between molecules so as to create segment combinations which could continue rapid and accurate replication could have been advantageous. Thus, it can be envisaged that a fourth need, the need to recombine, would have been another property which successful replicators would have optimized at an early stage.

If Tomizawa’s stem-loop “kissing” model is applicable to the early RNA world, then it can be further envisaged that replicators which modified their sequences to increase the probability of stem-loop formation (and hence of recombination), would have had a survival advantage. To reap the benefits of efficient recombination a replicator would have had to exchange segments with its own kind of replicator, not with other kinds. Thus need 4 implies a fifth need, the need to distinguish “self” from “not-self.” Finally, when the competition to build better “survival machines” began, a sixth need arose. This was the need to dedicate parts of the genome to the encoding of specific functions. This discussion deals with potential conflicts between needs.

![FIG. 5.—Linear regression analysis of the FORS-D value against the degree of exon overlap (%) for A, G0S19-1; B, c-fos; and C, p53. Each data point represents values for a 200-nt window. Each window overlaps its neighbors by 150 nt. For G0S19-1, windows extend from the beginning of the first exon to the end of the last exon. Windows in the flanks are omitted except those which are close to, and hence partially overlap, the first and last exons. For reasons given in the text, terminal exons are omitted from the analysis of c-fos and p53.](image-url)
### Table 1

Summary of FORS-D Plots and of Linear Regression Analyses of FORS-D versus Exon Overlap

<table>
<thead>
<tr>
<th>Gene</th>
<th>GENBANK NAME</th>
<th>LENGTH STUDIED (nt)</th>
<th>EXONS WITH DECREASED FORS-D VALUES</th>
<th>REGRESSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oncogene p53</td>
<td>HSP53G</td>
<td>8,750</td>
<td>- (+) (-) (+) (+) (+) (+) (-) (+)</td>
<td>-0.054*</td>
</tr>
<tr>
<td>Triose phosphate isomerase</td>
<td>HSTPI1G</td>
<td>4,400</td>
<td>- (+) (+) (+) (-) (+) (+) (-)</td>
<td>-0.011</td>
</tr>
<tr>
<td>Albumin</td>
<td>HUMALBGC</td>
<td>8,400</td>
<td>(-) (-) (-) - - - + (+)</td>
<td>-0.026</td>
</tr>
<tr>
<td>a-B-crystallin</td>
<td>HUMCRYABA</td>
<td>4,200</td>
<td>(-) (+) (+) (+) (+) (-)</td>
<td>+0.009</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>HUMCYCAA</td>
<td>3,088</td>
<td>- (-) (+) + - + (-)</td>
<td>-0.033*</td>
</tr>
<tr>
<td>Oncogene c-fos</td>
<td>HUMFOS</td>
<td>6,210</td>
<td>- - - (-) (-) (-) (-) (+)</td>
<td>-0.052*</td>
</tr>
<tr>
<td>Cytokine G0S19-1</td>
<td>HUMG0S19A</td>
<td>4,100</td>
<td>+ - (-) (-) (+) (+) (-)</td>
<td>0.041</td>
</tr>
<tr>
<td>Zinc finger protein</td>
<td>HUMG0S24B</td>
<td>3,135</td>
<td>- (+) (+) (-) (+) (+) (-)</td>
<td>-0.056*</td>
</tr>
<tr>
<td>β-globin</td>
<td>HUMHB</td>
<td>3,100</td>
<td>(-) (-) (-) (+) (+) (+) (-) (+)</td>
<td>+0.008</td>
</tr>
<tr>
<td>Heat shock protein 86A</td>
<td>HUMHSP86A</td>
<td>2,598</td>
<td>+ - (+) (+) (+) (+) (+) (-) (+)</td>
<td>-0.007</td>
</tr>
<tr>
<td>Heat shock protein 90B</td>
<td>HUMHSP90B</td>
<td>8,200</td>
<td>(+) (-) (-) (+) (+) (+) (-) (+)</td>
<td>-0.038</td>
</tr>
<tr>
<td>Trpuplin-C</td>
<td>HUMTROC</td>
<td>4,400</td>
<td>+ - (-) (+) (+) (+) (+) (-) (+)</td>
<td>-0.135</td>
</tr>
<tr>
<td>Tcp 10</td>
<td>MUSTCP10AA</td>
<td>3,882</td>
<td>- (-) (+) (+) (+) (+) (+) (+)</td>
<td>-0.066</td>
</tr>
<tr>
<td>Glucose regulated protein</td>
<td>TOMBIPGRP</td>
<td>2,966</td>
<td>(-) (+) (+) (+) (+) (+) (+) (+)</td>
<td>+0.017</td>
</tr>
</tbody>
</table>

*Exons showing relative decreases in FORS-D values (exemplified by figs. 3-4) are indicated by minus signs (−). Exons failing to show decreases in FORS-D values are indicated by plus signs (+). Parentheses around signs indicate some ambiguity.

**FORS-D values were plotted against degree of exon overlap (exemplified by fig. 5). P values indicate the extent to which slopes are not significantly different from zero.

c All exons were studied except HUMALBGC (first six only) and HSP53G. HUMHSP86A, and MUSTCP10AA (not first exon).

d Values with asterisks indicate that the last exon was omitted from the analysis.
Table 2
Comparison of Average FORS-D Values of Genes and Their Corresponding cDNAs

<table>
<thead>
<tr>
<th>GENE NAME</th>
<th>FORS-D (kcal/mol)</th>
<th>Genes</th>
<th>cDNA</th>
<th>DIFFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oncogene <em>p53</em></td>
<td>4.8 ± 0.4 (172)</td>
<td>4.2 ± 0.7 (47)</td>
<td>+0.6</td>
<td></td>
</tr>
<tr>
<td>Triose phosphate isomerase</td>
<td>3.7 ± 0.6 (85)</td>
<td>4.5 ± 0.7 (22)</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td>2.3 ± 0.4 (165)</td>
<td>0.4 ± 1.1 (12)</td>
<td>+1.9</td>
<td></td>
</tr>
<tr>
<td>α-B-crystallin</td>
<td>3.3 ± 0.5 (81)</td>
<td>1.8 ± 1.0 (10)</td>
<td>+1.5</td>
<td></td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>1.5 ± 0.3 (59)</td>
<td>2.2 ± 1.0 (22)</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>Oncogene <em>c-fos</em></td>
<td>1.9 ± 0.5 (120)</td>
<td>3.0 ± 0.8 (39)</td>
<td>-1.1</td>
<td></td>
</tr>
<tr>
<td>Cytokine <em>G0S19-1</em></td>
<td>4.2 ± 0.6 (79)</td>
<td>1.8 ± 1.2 (13)</td>
<td>+2.4</td>
<td></td>
</tr>
<tr>
<td>β-globin</td>
<td>2.7 ± 0.5 (59)</td>
<td>1.0 ± 1.2 (10)</td>
<td>+1.7</td>
<td></td>
</tr>
<tr>
<td>Heat shock protein 86A</td>
<td>2.2 ± 0.6 (49)</td>
<td>2.5 ± 1.1 (17)</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Heat shock protein 90B</td>
<td>2.2 ± 0.4 (161)</td>
<td>0.9 ± 0.7 (48)</td>
<td>+1.1</td>
<td></td>
</tr>
<tr>
<td>Troponin-C</td>
<td>4.4 ± 0.7 (85)</td>
<td>2.3 ± 1.9 (11)</td>
<td>+2.1</td>
<td></td>
</tr>
<tr>
<td>Mouse <em>Tcp-10</em></td>
<td>3.4 ± 0.6 (75)</td>
<td>2.6 ± 1.0 (24)</td>
<td>+0.8</td>
<td></td>
</tr>
<tr>
<td>Tomato glucose reg. protein</td>
<td>2.3 ± 0.5 (56)</td>
<td>2.2 ± 0.6 (22)</td>
<td>+0.1</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—The average FORS-D values (± standard error) for 200-nt windows were calculated for DNA segments containing the genes shown in table 1 and for their corresponding cDNAs. Numbers of windows, each overlapping the preceding window by 150 nt, are shown in parentheses. All genes are human except the bottom two.

Evidence for a Stem-Loop Model

It can be imagined that as part of the attempt to improve replicator stability (need 3), RNA molecules would have been superseded by DNA molecules, but these would have retained the potential to form stem loops. This leads to the prediction explored here that, where not constrained by other needs, modern DNA molecules would have a special potential to form stem-loop structures. This would be particularly evident during meiosis (Kleckner and Weiner 1993) and might require an exceptional degree of supercoiling for DNA to depart from its classical duplex structure (Murchie et al. 1992). Consistent with this, inhibitors of topoisomerases, which relax supercoiled DNA, greatly enhance recombination (Wang et al. 1990). Furthermore, endonucleases involved in recombination, which make single-strand nicks in double-stranded DNA, require a supercoiled substrate (Sung et al. 1993). Although a relatively crude measure, there is evidence from studies with anticrcuciform antibodies for the presence of stem-loop structures, which can vary with the physiological state of the cell (Ward et al. 1990). Further evidence for a role of stem loops in recombination is presented elsewhere (Romanova et al. 1986; Reeder et al. 1994).

Positive FORS-D values indicate the existence of an evolutionary pressure on DNA base order, rather than on base composition, to generate the potential to form stem-loop structures. Thus, in the absence of other constraints, base changes which enhance the formation of stem loops would have been accepted. I have reported here that dispersed positive FORS-D values are found in long human genome segments considered to consist mainly of nontranscribed intergenic DNA (fig. 1). This is found with DNA from various species, including some with much less intergenic DNA than humans (Forsdyke 1995b). The most obvious genome-wide selective force driving evolution of the potential to form stem loops is recombination.

Conflict Hypothesis and the Evolution of Introns

A potential conflict between the genome-wide FORS-D pressure and pressures on DNA to encode specific functions is apparent (fig. 2). A sequence required to encode a protein might not at the same time be able locally to optimize its folding propensity. The conflict can be met in three ways. First, because of the redundancy of the genetic code, particular synonymous codons can be preferred. Second, amino acids with similar functions (e.g., serine and threonine) can be interchanged to widen the range of codon choices. Third, the sequences encoding a protein can be diffused over a wide region by permitting encoding to occur only in discrete segments. If the first two options are not sufficient, then only the third option is left. Thus introns might correspond to parts of a gene where the constraints on the first two options are most severe. The more severe the constraints, the more introns there would be, and the longer would be the length of DNA occupied by the gene. This conflict hypothesis might explain how introns
initially arose. However, it does not appear to address the problem of how introns came to vary so dramatically in length (Hawkins 1988). Variation in intron length would tend to prevent recombination between genes with homologous exon sequences by impairing the precise register required for successful "kissing" between the loops of stem-loop structures.

In terms of FORS-D values, the hypothesis states that when exons first arose, the exon sequences themselves would have had low FORS-D values, except where, by chance, the need to encode a peptide happened not to be in conflict with the need to form stem loops. The low FORS-D exons would have been surrounded by high FORS-D introns and flanking sequences. Over evolutionary time the demands of regulation of gene expression would have been imposed upon this primitive arrangement, so that low FORS-D values might be present in regions of DNA where regulatory proteins bind (except, perhaps, where the regulatory proteins recognize DNA palindromes). Also, over evolutionary time a background "noise" might have been imposed by mobile genetic elements and recombinations accompanying exon shuffling (Gilbert and Glynias 1993). Taking into account the proposed long evolutionary period since introns were established in genes, the analyses of FORS-D values in various intron-containing genes (figs. 3–4; table 1) do provide some support for the idea of a conflict between genome-wide and local pressures. Further evidence has been provided by recent studies of snake venom phospholipase A₂ genes, retroviral genomes, and major histocompatibility complex peptide-binding genes (Forsdyke 1995a, and unpublished data). All of these are under greater selection pressure (positive selection) than the genes studied here (with the possible exception of genes encoding troponin C; Ohta 1994).

Some terminal long open-reading frames were excluded from the linear regression analyses shown in figure 5 and table 1. In terms of the hypothesis presented here, long open-reading frames would only exist in certain genes because those genes had been able to resolve the conflict in the regions where we now see the long open-reading frames; FORS-D pressure would have been accommodated by the choice of appropriate synonymous codons or of amino acids with similar functions. (These would have been the main choices open to organisms without introns.) Thus, FORS-D analyses of genes with long open-reading frames are not inconsistent with the conflict hypothesis (fig. 4). By omitting unduly long exons from regression analyses, one hopes to better assess genomic regions which are less able to adapt to FORS-D pressure by using synonymous codons or functionally similar amino acids.

The conflict hypothesis would predict that cDNA sequences would tend to have lower average FORS-D values than the genes from which they derive. However, it does not follow that low FORS-D exons, when united, will always result in a low FORS-D cDNA. Two independently arising exons may each have a low potential to form stem loops but a high potential to form stem loops collectively. As a simple example, consider two exons of sequence CACACACA and TGTGTGTG. Both have low FORS-D values, but they would unite to generate a cDNA with a very stable stem loop (high FORS-D value). Only some genes show clear evidence for higher average FORS-D values than the corresponding cDNAs (G0S19-I, fig. 3; table 2).

Why Are FORS-D Values So Sensitive?

Evolutionary pressures (mutation, selection, drift) determine how many bases there are in a sequence (DNA length), the proportions of different bases (base composition), and the order of the bases (sequence). The experimental shuffling of the order of bases in a sequence of given length and base composition disrupts information present in the natural sequence and generates a reference sequence of the same length and base composition against which the natural sequence can be compared. Thus one can distinguish evolutionary pressures affecting base order (primary sequence) from those affecting base composition. FORS-M and FONS profiles closely follow each other (figs. 1B, 3B–4B), showing that base composition is the major determinant of the stability of stem loops. It thus might seem somewhat surprising that differences of only a few kilocalories (FORS-D values) provide such a sensitive indicator. That fluctuations in FORS-D values are not random is shown by their consistent average positivity (fig. 1) and the association of negative FORS-D values with functionally important regions of DNA (fig. 3).

Base composition is a genome "strategy" (Grantham et al. 1980) and thus is less likely to be relevant to local gene function than primary sequence. FORS-D values provide a sensitive indicator because they measure the base order–determined stem-loop potential, not the base composition–determined stem-loop potential. A close examination of folding energy profiles reveals that high FORS-D values may sometimes be associated with quite low negative FONS values (a function of base order and base composition) and even lower negative FORS-M values (a function of base composition). Thus, high negative FONS values (implying a very stable secondary structure) may sometimes have little local functional relevance.

The recombination model of Wagner and Radman (1975) proposes that, after initial contact is made by loops at the tips of the stems, there are endonucleolytic cleavages of single strands of opposing stems. Single-strand exchanges between the stems then follows. This
process would presumably abort if precise complementarity were not present in the stems (Radman and Wagner 1993). Furthermore, Tomizawa (1993) has concluded that the major role of the stem in a stem-loop structure is the proper positioning of the loop. This allows the unpaired bases in the loop to pair, in register, with those of an appropriately positioned loop projecting from a complementary nucleic acid. This "kissing" is rate limiting in recombination. Thus, a departure, even by a few kilocalories, in complementarity between bases in the stem might prevent recombination. An evolutionary force adapting base order to favor stem-loop formation could be a very powerful one.

Introns Early or Late?

In that the ability to recombine is held to have evolved before the ability to encode proteins, this paper supports the "introns early" model. The positions of exon-intron junctions are held to have been determined by the need to form stem loops and are not necessarily related to protein domains (Gilbert and Glynias 1993). Stoltzfus et al. (1994, p. 202) concluded that "no significant correspondence between exons and units of protein structure was detected."

GC/AT Fine-Tuning and Speciation

Different genomes, or genome compartments, have characteristic (G+C)/(A+T) ratios (Wyatt 1952; Bernardi 1989; Filipski 1990). In vertebrates this "GC/AT pressure" (fig. 2) is evident in intergenic regions, and introns, and in both coding and noncoding parts of exons. The "genome hypothesis" (Grantham et al. 1980) has led to the proposal that there is some intrinsic genome-wide pressure favoring the adoption of a particular (G+C)/(A+T) ratio in a particular species. The view that this primarily reflects mutational biases (Filipski 1990) has been challenged (Bernardi 1989). I argue elsewhere (unpublished data), that species-specific settings of the (G+C)/(A+T) ratio have arisen as part of a fine-tuning process which prevents recombination between species. Just as different radio transmitters broadcast their messages at different wavelengths to avoid interference, so different genomes "broadcast" their sequences at different (G+C)/(A+T) "wavelengths," to avoid undesirable recombination events.

The present work suggests a mechanism by which this could occur. Base composition, rather than base order, is the major factor determining the folding energy of a nucleic acid segment (figs. 1B, 3B-4B). Thus small changes in base composition could greatly affect the looping pattern which a sequence could present for homology search. It would be more difficult to extrude loops from GC-rich DNA than from AT-rich DNA. Two sequences of different G+C percentages undergoing supercoiling in a common intracellular environment might extrude stem loops at different times and to different extents. The pattern of loops presented by homologs would be different, and impaired recombination would result. To recombine, two sequences must be equal both in the local parameter (base order-determined stem-loop potential) and in the genomic parameter (base composition-determined stem-loop potential).

If avoidance of recombination between different species (including a host species and its pathogen species) is evolutionarily advantageous (Bernstein and Bernstein 1991), then there should have been a selection pressure such that each species in a common environment with other species would have fine-tuned its (G+C)/(A+T) ratio to a level distinct from those of the other species. Thus, biologically similar species may have dissimilar ratios, whereas biologically dissimilar species (unlikely to interact sexually and with dissimilar pathogens) may have similar ratios (Wyatt 1952). The fine-tuning of base ratios could have been a key component of the postzygotic isolation process leading to speciation (unpublished data).

Acknowledgments

I thank D. Back for assistance in computer configuration and M. Go, C. Tittiger, V. K. Walker, and G. R. Wyatt for helpful comments on the manuscript. The work was supported by a grant from the Medical Research Council of Canada.

LITERATURE CITED


