Chapter 10

Exons and Introns

“Ideas in the theory of evolution can be used in situations far removed from biology. Similarly, information theory has ideas that are widely applicable to situations remote from its original inspiration.”

Richard Hamming (1980) [1]

Organisms can be divided into those whose cells do not have a nucleus, the single celled “prokaryotes” (Greek: pro = before; karyon = nucleus), and those whose cells have a nucleus, the single- or multi-celled “eukaryotes” (Greek: eu = good or normal; karyon = nucleus). Prokaryotes include species of bacteria (eubacteria) and archaea (archaebacteria), the latter being a bacteria-like group sometimes found in extreme environments (e.g. hot springs). Eukaryotes (eukarya) include all species of animals and plants, both single-celled (protozoa, protophyta) and multi-celled (metazoa, metaphyta).

The etymology may suggest that organisms without a nucleus, somewhat like modern bacteria or archaea, evolved before organisms with a nucleus. However, as discussed in Chapter 8 when wondering whether isochores appeared early or late, prokaryotes and eukaryotes are believed to have evolved from a common ancestor, whose properties we can now only infer. It may appear parsimonious to imagine an ancestor without a nucleus. But parsimony may be in the mind of the beholder. Unlikely maybe, but perhaps the ancestor had a nucleus?

In the 1960s many eukaryotic RNAs were found to be first synthesized as “giant” RNAs, a large part of which was subsequently “wasted” as the RNAs were “processed” to generate mature cytoplasmic products. Considering a “gene” as corresponding to a unit of transcription, it appeared that a large part of a gene might contain redundant information. Eukaryotic ribosomal RNAs (rRNAs) were transcribed from DNA as long precursor RNAs that were subsequently processed by the removal of apparently functionless internal “spacer” sequences [2]. However, prokaryotic (bacterial) rRNA genes were more compactly organized. Did rRNA genes in the first organisms to
evolve have the spacer sequences, which decreased later in prokaryotes ("spacers-early"), or were the long spacer sequences acquired later in eukaryotes ("spacers-late")?

A similar processing was found in the case of eukaryotic messenger RNAs (mRNAs). Again, bacterial mRNA genes were more compactly organized. By 1970 it had been shown that many freshly minted giant nuclear RNAs contained sequences corresponding to smaller mature cytoplasmic mRNAs. A simple model was that each giant RNA consisted of an mRNA sequence that was flanked by long redundant RNA segments [3]. It turned out, however, that instead of one DNA sequence segment compactly encoding one protein, as in most prokaryotic genes, eukaryotic genes often encode proteins in sequence segments that are interrupted by sequence segments that do not usually encode proteins. Thus, if a mature "sense" mRNA be represented as:

\[
\text{Mary had a little lamb, its fleece was white as snow} \quad (10.1)
\]

Then the corresponding giant precursor RNA would contain segments of, what would appear to be, "nonsense:"

\[
\text{Mary had a qvhxmg eqz d xgw tepcrmbits fleece was white as snow} \quad (10.2)
\]

Since an entire gene was transcribed, the internal RNA sequences (e.g. xqvhxmg eqz) derived from DNA "introns," had to be removed from the initial transcript. What remained in the processed RNA, the mRNA, was derived from DNA "exons" (Fig. 10-1) [4]. Thus, a protein-encoding gene consists of exons and introns. Since the "spacer" phenomenon had already been described for rRNA it should not have been a surprise that it also applied to mRNAs; but many, myself included, were surprised at its generality.

This chapter presents evidence that introns are a way of resolving intragenomic conflicts between different forms and levels of information. Cases of extreme intragenomic conflict, as when genes in "arms races" evolve rapidly under positive selection, demonstrate most clearly that "information theory has ideas that are widely applicable" [1].

**Introns Interrupt Information**

Although there were many attempts, it proved difficult consistently to associate exons with domains of protein structure or function [5, 6]. Introns interrupt genetic information per se, not just protein-encoding information. Thus, intron-encoded information is removed during transcript processing both from parts of mRNA precursors that encode protein, and from parts of
mRNA precursors that do not encode protein (i.e. from 5' and 3' non-coding sequences). Certain special RNAs which, like rRNAs, are not translated to give a protein product (e.g. see Chapter 14), also have introns.

Fig. 10-1. Transcription and processing of mRNA from a eukaryotic gene. Sequences that correspond to the final mRNA are marked as boxes on the duplex DNA. The enzyme RNA polymerase initiates transcription at the left end of exon 1 using the bottom DNA strand as template, and terminates transcription at the right end of exon 3. Intron sequences are spliced out from the initial transcript, so that the mature mRNA consists of a 5' non-protein-coding region, a central protein-encoding region, and a 3' non-protein-coding region. The central region is translated into protein. A gene may be narrowly construed as corresponding to the segment of DNA that is transcribed. A broader definition would include other segments (e.g. the promoter region where RNA polymerase initially binds to DNA), which may lie outside the transcribed segment. Williams gives a different definition (see Chapter 8).

Exons have a narrow size range, with a peak at about 100 bases (Fig. 10-2) [7]. If coding, this would correspond to 33 amino acids. So proteins, often containing hundreds of amino acids, are usually encoded by long genes with many exons [8]. There appears to be a limit to the extent to which a genome will tolerate a region that both encodes protein, and is purine-loaded (see Chapter 6). When that limit is reached, protein-encoding and purine-loading functions are arrested, and can restart only after a "decent" interval – namely
an intron interval. Introns are generally longer than the exons that surround them (see Fig. 2-5). Each exon being separated from its neighbours by introns, the sum of intron lengths tends to increase proportionately as the sum of exon lengths increases [9].

![Distribution of exon sizes in eukaryotic genes.](image)

**Fig. 10-2.** Distribution of exon sizes in eukaryotic genes. The number of exons in different size categories increases to a maximum and then progressively declines. Most exons are about 100 bases in length. (Data are from the ExInt database)

What function(s), if any, do introns have? Did introns appear early or late in the evolution of biological forms? If introns can be dispensed with in bacteria, then perhaps they have no function. Alternatively, whatever function introns have, either is not necessary in bacteria, or can be achieved in other ways by bacteria. Since members of many bacterial species appear to be under intense pressure to streamline their genomes to facilitate rapid replication, if it were possible they should have dispensed with any preexisting introns and/or should have been reluctant to acquire them. On the other hand, if introns play a role, and/or do not present too great a selective burden, then organisms under less pressure for genome compaction might have retained pre-existing introns, or might have acquired them.

An early origin of introns is suggested by the fact that the humans share the positions and sequences of many introns with a marine worm, indicating that their common ancestor had the same introns [10]. Introns did not first appear so that at some remote future date organisms with exons would be advantaged. Nature is not prescient. Although something playing a role at one point in time can come later to serve a quite unrelated role, in general evolu-
tion does not work this way. Sometimes a random event (genetic drift) provides an evolutionary toe-hold, but for something so widespread and drastic as introns there should be some immediate selective advantage.

Knowing the function of introns seemed critical for sorting this out. There were many ingenious suggestions. Some thought introns were just another example of the apparently useless “junk” DNA that appeared to litter the DNA of many organisms (see Chapter 12). Others thought that introns might have facilitated the swapping of protein domains to generate new proteins [4, 11], but that did not explain how introns initially arrived.

However, the notion of “message” sequences interrupted by “non-message” sequences is familiar to those working on noise affecting signal transmission in electrical systems. In these systems information scientist Richard Hamming pointed out that the non-message sequences can have an error-checking function that permits the receiver to detect and correct errors in the message sequence [1]. Could introns have a similar error-checking role [12]?

It appears that the order of bases in nucleic acids has been under evolutionary pressure to develop the potential to form stem-loop structures, which might facilitate “in-series” or “in-parallel” error-correction by recombination (see Chapter 2). This means that genomic sequences convey more than one level of information (see Chapter 9). Furthermore, as predicted in 1893 by the discoverer of DNA, Johann Miescher, a sexual process that brings molecules from separate sources together, could facilitate the mutual correction of errors (see Chapters 3 and 14). However, the need to participate in the process of error-detection and correction can result in redundancies and various constraints (see Chapter 4).

The error detection and correction process requires an alignment of two sequences, which itself depends on an initial “homology search.” Thus, there must be sufficient similarity between two sequences for a successful homology search. One outcome of this is that segments of DNA link up (“recombine”) with other segments. In the course of this swapping of segments, errors can be detected and corrected (“gene conversion;” Fig. 8-3). The process is referred to as recombination repair, which distinguishes it from a variety of other repair processes that will not be considered here [13]. The adaptive value of recombination repair is likely to be very great. So, if it could have arisen early in the evolution of primeval biological forms (perhaps in an “RNA world”) prior to the evolution of protein-encoding capacity (i.e. if it were chemically feasible), then it would have arisen. Williams noted in 1966 [14]:

“The existence of genetic recombination among the bacteria and viruses, and among all of the major groups of higher organisms, indicates that the molecular basis of sexuality is an ancient evo-
volutionary development. Our understanding of the structure of the DNA molecule makes recombination at this level easy to visualize. In a sense sex is at least as ancient as DNA. ... I would agree, therefore, ... that sexual reproduction is as old as life, in that the most primitive living systems were capable of fusion and of combination and recombination of their autocatalytic particles. Modern organisms have evolved elaborate mechanisms for regulating this primitive power of recombination and for maximizing the benefits to be derived from it.”

Given an early evolution of recombination, protein-encoding capacity could then have had to intrude into the genomes of biological forms already adapted for recombination. Would this intrusion have been readily accepted? Or would protein-encoding capacity have had to elbow its way forcibly into primeval genomes?

**Protein Versus DNA**

Although the degeneracy of the genetic code provides some flexibility as to which base occupies a particular position, there may still be a conflict between the needs of a sequence both to encode a protein and to respond to other pressures. Situations where protein-encoding and/or other pressures are extreme should be particularly informative in this respect.

Extreme protein-encoding pressure is apparent in the case of genes under very strong positive Darwinian selection. In Chapter 7 competition among speakers for the attention of an audience provided a simple metaphor for positive selection. In that case, speakers were positively selected if they could communicate rapidly to a fixed audience by overcoming idiosyncrasies of accent. Under biological conditions, however, “the audience” is not fixed. Positive selection often occurs under conditions where both “speakers” and “audience” are rapidly changing. This includes genes affected by “arms races” between predators and their prey.

For example, snake venom may decrease the rodent population (prey) until a venom-resistant rodent line develops and expands (i.e. a mutant line arises with this selective advantage). Now, while the rodent population expands, the snake population (predators) decreases because its members cannot obtain sufficient food (i.e. rodents). This decrease continues until a line of snakes with more active venom, which can overcome the resistance, develops and expands (i.e. a mutant line arises with this selective advantage). This population of snakes now expands, and the rodent population begins to fall again.

The cycle constitutes a biological arms race, and influences particular gene products. Parts of venom proteins which are important for toxicity are required to change so rapidly in response to this strong pressure from the envi-
ronment (i.e. from rodents), that the corresponding genes can no longer afford the luxury of both encoding the best proteins (primary information) and attending to other pressures (secondary information). They must encode better proteins even at the expense of their abilities to respond to other pressures. Accordingly, under extreme positive selection pressure the rate of sequence change in protein-encoding regions is high (Fig. 10-3).

**Fig. 10-3.** High base substitution frequency and low base order-dependent stem-loop potential in exons of the rattlesnake gene encoding the basic subunit of venom phospholipase A\(_2\) (PLA\(_2\)), which is under positive Darwinian natural selection. The distribution of base substitutions (continuous line in (a)) is compared with values for FORS-D (closed triangles in (a)), FORS-M (open triangles in (b)), and FONS (closed circles in (b)). Values were determined for overlapping 200 base windows, which were moved in steps of 50 bases. Substitutions are base differences relative to the rattlesnake PLA\(_2\) acidic subunit gene. The two genes are likely to have arisen by duplication of a common ancestral gene. Boxes in (b) indicate the location of the four exons, with dashed vertical lines showing, consecutively, the beginning of exon 1, the beginning of the protein-coding part of exon 1, the end of the protein-coding part of exon 4, and the end of exon 4.
When sequences of similar venom proteins (e.g., phospholipase A2) from two snake species (or from duplicated genes within a species) are compared, great differences are found in the protein-encoding parts of exons (i.e., low sequence conservation). In contrast, small differences are found in introns, and in the 5' non-coding and 3' non-coding parts of exons (i.e., high sequence conservation; Fig. 10-3a). This is a dramatic reversal of the more usual situation where, in genes under classical negative ("purifying") Darwinian selection, exons are conserved much more than introns (i.e., introns display more variation; see Fig. 2-5).

What is being conserved in snake venom introns? Analysis of fold potential as it affects base order (FORS-D; see Chapter 5) reveals that base order-dependent stem-loop potential is low in exons (where sequence conservation is low) and high in introns (where sequence conservation is high; Fig. 10-3a). Base order-dependent stem-loop potential appears to have been conserved in introns; indeed, there is an inverse (reciprocal) relationship between base substitution frequency and base order-dependent stem-loop potential. When base order-dependent stem-loop potential is high, base substitution frequency is usually low (i.e., sequence conservation is high).

This suggests that the pressure to adapt the protein sequence (requiring non-synonymous codon changes) has been so powerful that base order has not been able to support base order-dependent stem-loop potential in the same exon sequence (Fig. 5-6). Instead, stem-loop potential is diverted to introns, which are appropriately conserved (fewer base substitutions than the surrounding exons). This is in keeping with the hypothesis that early in evolution protein-encoding potential was imposed on prototypic genomes that had already developed stem-loop potential. For this imposition to succeed without disturbing the general distribution of stem-loop potential, proteins had to be encoded in the fragments that we now call exons [15-17]. Thus, in the general case, introns were "early."

Another example of positive selection is the genome of the AIDS virus (HIV-1), which can be viewed as a predator, with us as its prey (see Chapter 8). Here, an inverse correlation between substitutions and base order-dependent stem-loop potential can be observed when the disposition of substitutions and fold potential along the genome are displayed (Fig. 10-4). At first glance the data appear as a confused jumble of lines. But when the paired values from along the sequence are plotted against each other, a significant inverse correlation emerges (Fig. 10-5c) [18, 19].

Thus, sequences varying rapidly in response to powerful environmental selective forces ("arms races") appear unable to order their bases to favor the elaboration of higher order folded structures (of a type that, in eukaryotes, might mediate meiotic chromosomal interactions and recombination repair; see Chapter 8). So, the encoding of nucleic acid stem-loop structure can be
relegated either to non-protein-encoding regions, namely, introns, 5' and 3' non-coding regions, and non-genic DNA (the favored option in less compact genomes), or to 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).

Fig. 10-4. High base substitution frequency and low base order-dependent stem-loop potential in regions of the AIDS virus genome that are under positive Darwinian natural selection. The various genes are shown as grey boxes, with their abbreviated names attached. Two genes, TAT and REV, each have two exons that are shown linked by continuous lines that represent introns. Thus, one gene's intron can be another gene's exon. The distribution of base substitutions for virus "subtype" HIVSF2 relative to virus "subtype" HIVHXB2 is shown as a continuous line. Values for base order-dependent stem-loop potential (FORS-D) are shown as filled triangles. All values are for 200 base overlapping windows, which were moved in steps of 50 bases.

Functionally important regions are conserved in genes evolving slowly under classical negative selection (i.e. there is a low local base substitution frequency in protein regions under this selection pressure). If the rate of evolution has been slow, then there has been more time to arrive at an appropriate compromise with base order-dependent stem-loop potential. Thus, in the case of slowly evolving sequences, a relationship between base order-dependent stem-loop potential and sequence variability may be less evident. On the other hand, the demands of faithful reproduction of a protein, with negative selection of individuals bearing mutations affecting its functionally
most important parts, can leave the co-encoding of stem-loops not only to non-protein-encoding regions (e.g. introns), but also to regions encoding functionally less important, and hence more variable, parts of proteins (such as the protein surface, which cell water can readily access) [20, 21]. In this case, high base order-dependent stem-loop potential can correlate positively (not inversely) with high substitution rates (variability) when similar (homologous) sequences from different species are compared. Indeed, comparison of certain human and mouse oncogenes (FOSB) reveals a positive correlation between base order-dependent fold potential and substitution frequency [22].

Fig. 10-5. In the AIDS virus genome there is (c) an inverse relationship between base order-dependent potential (FORS-D values) and substitution frequency, but (b) no detectable relationship between base composition-dependent stem-loop potential (FORS-M) and substitution frequency. Shown in (a) are FONS values, which are the sum of FORS-D and FORS-M values. Although only approximately 10% of the variation in substitutions can be explained by base order-dependent stem-loop potential (shown by the $r^2$ value in (c)), the downward slope is significantly different from zero (shown by the low $P$ value in (c))
**Achilles Heels**

The AIDS virus causes millions of deaths annually, and subtypes resistant to therapeutic agents have become more prevalent. Given the profiles shown in Figure 10-4, can you suggest potential targets for therapeutic attack—namely targets that are least likely to differ between different pathogens? Note that the region to the left of the GAG gene in the AIDS virus has the lowest base substitution rate (i.e., it is the most conserved part of the genome). This coincides with a major peak in base order-dependent stem-loop potential (indicated by high negative FORS-D values). This suggests that there is here a special need to conserve stem-loop potential [18]. Indeed, this is the location of the “dimer initiation” stem-loop sequence, which is necessary for the copackaging of two AIDS virus genomes as part of a process that resembles meiotic pairing (see Chapter 8). This may yet prove to be the Achilles heel of the AIDS virus, once the problem of “flushing out” latent forms from host genomic reservoirs is solved [23, 24].

We humans may also have an Achilles heel. Our genomes are rich in palindromes conferring stem-loop potential (Fig. 5-2), which is usually to our advantage (see Chapters 8 and 14). But sometimes the palindromes involve oligonucleotide repeats, which can be disadvantageous, as will be discussed in Chapter 11 (Fig. 11-8). This is particularly so when the repeats are AT-rich and consequently can readily adopt cruciform stem-loop configurations.

Palindromic regions containing AT-rich repeats are prone to recombine, at times when homologous chromosomes may not be precisely aligned, by kissing interactions with other palindromic regions containing AT-rich repeats. Thus, there may be “cut-and-paste” translocations (transpositions) between non-homologous chromosomes (“illegitimate recombination”), which can be detected in normal sperm samples [25]. Whether pathological results occur depends partly on the frequency of the translocations. This, in turn, depends on the lengths of the regions containing AT-rich repeats, which varies among individuals (polymorphism). It is likely that, by virtue of this Achilles heel, many individuals with long palindromic regions containing AT-rich repeats have been eliminated by natural selection.

**Mirror Repeats**

In Chapter 4 we encountered “inverted repeats” with palindromic properties at the duplex level (see also Fig. 2-4). For example a “top” single-strand in a duplex might read:

\[ 5' \text{AAAAACCCCGGGTTTTT} 3' \]  

(10.3)
Here \textit{AAAAACCC} in the 5' half of the top strand is repeated on the complementary strand, where it pairs with the \textit{GGGTTTTT} sequence in the 3' half of the top strand. Such sequences appear to serve a DNA level function, since they facilitate the extrusion of stem-loop secondary structures (Fig. 5-2). Single-strands also contain "direct repeats" that might, for example, read:

\begin{equation}
5' \text{AAAAACCC} \text{AAAAACCC} 3' \quad (10.4)
\end{equation}

Here \textit{AAAAACCC} is repeated. Single-strands can also contain "mirror repeats" that might read:

\begin{equation}
5' \text{AAAAACCCCCCAAAAA} 3' \quad (10.5)
\end{equation}

Here \textit{AAAAACCC} can be considered the "mirror" of \textit{CCCCAAAA}. In these two cases, (10.4) and (10.5), the repeats occur in the same strand. Direct repeats and, especially, mirror repeats, have the potential to oppose local stem-loop formation, and so to oppose any DNA level function that stem-loops might serve.

Remarkably, mirror repeats are found at particular locations in exons. Their locations correlate with the boundaries of various structural elements in proteins; indeed, mirror repeats can \textit{predict} where such structural elements will occur [26]. In this case it appears that a conflict between protein and DNA has been won by protein. By preventing stem-loop formation mirror repeats should prevent local recombination and thus preserve the local integrity of the DNA encoding a structural element in a protein.

We know that a protein, by "insisting" (through natural selection) on having a particular amino acid at a particular position in its sequence, requires that the corresponding gene have a suitable codon at a particular position in its sequence. Now we see that a protein, by "insisting" (through natural selection) on having a particular structural element (e.g. alpha-helix, beta-strand), also requires that the corresponding gene have appropriately positioned mirror repeats (albeit often imperfect mirror repeats).

\section*{RNA Versus DNA}

Usually a particular DNA sequence is transcribed into an identical RNA sequence, with the exception that RNA molecules have \textit{U} (uracil) rather than \textit{T} (thymine); but these are chemically similar bases. So it is not surprising that, in broad features, computer-derived secondary structures for an RNA molecule (using dinucleotide pairing energy tables for RNA bases), are similar to the structures derived for the corresponding DNA (using dinucleotide
pairing energy tables for DNA bases; see Table 5-1). Yet there are genes with no protein product. The gene products are RNAs, which have specific functions dependent on the secondary (and higher order) structures they adopt (often selected for at the cytoplasmic level). If such stem-loop secondary structures also sufficed for function at the DNA level, then there might be no need for introns in genes for non-protein-encoding RNAs.

The fact that there are spacers or introns in such genes, implies that sequences generating the stem-loop secondary structures that suffice for function at the RNA level, do not suffice for, and may even conflict with, the sequences needed for stem-loop secondary structures that 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 cytoplasmic level; see Chapter 6), then purine-loading pressure (which would constrain stem-loop patterns in exons) should support stem-loop pressure in provoking the splitting of what might otherwise have been large exons.

### Overlapping Genes

Strict adherence to the RNY-rule (see Chapter 7) would dictate that, of the three possible triplet reading frames in the “top” mRNA synonymous “coding” strand of DNA, the one that best fits the RNY pattern would be the actual reading-frame. However, sometimes it is expedient for genes to overlap, either entirely or in part, and in this case one of the genes, if transcribed in the same direction, can use another, non-RNY, reading-frame. This applies to some of the genes of the AIDS virus, which are all transcribed to the right (Fig. 10-4).

In some circumstances, genes transcribed in different directions may overlap. Thus, the “top” strand may be the coding strand of one gene, and the “bottom” strand may be the coding strand of another gene. Again, one reading frame is RNY and two are non-RNY (Fig. 7-2), and any of the three may be employed.

Whatever the transcription direction, in overlapping genes the region of the overlap can come under extreme protein-encoding pressure, and this might conflict with other pressures. Indeed, consistent with the argument made here, base order-dependent folding potential is constrained where genes overlap [27].

### Simple Sequences

As in the above examples of genes under extreme protein-encoding pressure, genomes under another extreme pressure, GC-pressure, should also be highly informative. The genome of the most lethal malaria species, *Plasmodium falciparum*, satisfies this requirement, being under strong “downward
GC-pressure.” It is one of the most AT-rich species known (i.e. very low (G+C)%; see Fig. 9-5a). Another unusual feature of *P. falciparum* is that many proteins are longer than their equivalent proteins (homologues) in species that have less extreme genomic (G+C)% deviations. This reflects the acquisition by the *P. falciparum* proteins of low complexity “simple sequence” segments that have no known function. Simple sequence at the protein level (i.e. runs of amino acids from a limited range of the twenty possible amino acids) is encoded by simple sequence at the nucleic acid level (i.e. runs of bases from a limited range of the four possible bases; see Chapter 11).

There are other unusual features of the *P. falciparum* genome. Unlike many eukaryotic genomes, there is poor correlation between the length of a gene and the combined lengths of the introns of that gene (Fig. 10-6a). Yet there is a close correlation between the length of a gene and the combined lengths of low complexity segments in that gene (Figs. 10-6b). In this respect the low complexity elements appear like introns; however, unlike introns, they are not removed during processing of the RNA transcript.

Furthermore, introns and low complexity segments are not interchangeable in that, as absolute intron length increases, there is little decline in length of low complexity segments in a gene (Figs. 10-6c). It is only when the lengths are expressed as a proportion of gene length that a reciprocal relationship emerges (Figs. 10-6d). Whereas in many eukaryotic genomes intron locations show no relationship to protein functional domains (since if splicing is accurate intron location is irrelevant to the protein), low complexity segments in *P. falciparum* must, of necessity, predominate between functional domains.

If low complexity segments have no, or minimal, function at the protein level, do they reflect a function at the nucleic acid level? Low complexity segments in *P. falciparum* are usually of high (A+G)% and so they contribute to purine-loading. Introns, in contrast, tend to be of low (A+G)% (i.e. tending towards pyrimidine-loading). It will be shown in Chapter 11 that, when contributing to AG-pressure (purine-loading pressure), low complexity segments can countermand fold potential. But when AG-pressure is not extreme, low complexity segments do contribute to fold potential. In this respect, they do resemble introns.

Spacers and introns are likely to have arisen early in evolution because they are preferential sites for the encoding of the stem-loop structures in DNA that are necessary for initiating recombination and, hence, error-detection and correction. While, in extreme cases, by virtue of this function, introns are conserved more than exons (Fig. 10-3), in the general case, to facilitate the anti-recombination necessary for gene or genome duplication (speciation), introns evolve more rapidly (are conserved less) than exons (see Fig. 2-5 and Chapter 8).
Fig. 10-6. In the genome of the malaria parasite, *P. falciparum*, low complexity segment length, not intron length, correlates best with gene length. Shown for chromosome 2 are relationships of lengths of introns (a), and low complexity segments (b), to the lengths of the corresponding genes, and relationships of either absolute (c), or percentage (d), lengths of low complexity segments to the lengths of introns in the same genes. *P* values indicate the probability that the slope values are not significantly different from zero (i.e. a low *P* value indicates high significance).
Multiple Pressures

The multiple, potentially conflicting, pressures affecting both the genome phenotype and the conventional (classical Darwinian) phenotype are summarized in Figure 10-7 and Table 10-1.

**GC and Fold Pressures**

Fig. 10-7. Summary of potentially conflicting pressures operating at the mRNA level. The genome-wide pressures, GC-pressure and fold (stem-loop) pressure, influence the entire mRNA (shown as a thick horizontal arrow with loops enclosed in the large outer box). Purine-loading pressure (AG-pressure) is a local pressure that also influences the entire mRNA. Small grey boxes indicate potential sites for the binding of regulatory factors that sometimes preferentially locate to the 5' and 3' non-coding regions. The large grey box indicates the central, protein-encoding region, where "protein-pressure" is deemed to operate.

The proteins of *P. falciparum* provide an extreme example of nucleic acid level pressures that affect protein sequence. However, proteins of all organisms are, to a degree, manifestations of nucleic acid level pressures. The correlation between gene length and content of low complexity segment (Fig. 10-6b) is probably general, with low complexity segments seeming to serve pressures for purine-loading, rather than for protein function. In a species where purine-loading was a dominant nucleic acid level pressure, proteins should be longer (and hence generally larger) than their homologous proteins in other species where purine-loading was less dominant. Since there is a reciprocal relationship between GC% and purine-loading (Fig. 9-7), then proteins should tend to be smaller in GC-rich species (Fig. 9-9), unless such GC-
richness itself required addition of GC-rich low complexity segments. Since GC-rich species tend to have small proteins, it is likely that purine-loading is generally the dominant nucleic acid pressure in this respect (see Chapter 11).

| Environmental selective factors | Selection for mutations which: | Primary effect on DNA function | Biological result | Observed features of modern DNA |
|---------------------------------|---------------------------------|---------------------------------|------------------|--------------------------------|
| Classical phenotypic selective factors | Change encoded proteins | Protein-encoding can constrain other functions | Change in classical phenotype | Changed base-pair in accordance with Chargaff's first parity rule |
| Competitors with more efficient translation, and intracellular pathogens | Purine-load RNAs | Purine-loading can constrain other functions | Efficient translation, and no "self" double strand RNA formation | Chargaff's cluster rule and Szybalski's transcription direction rule |
| Mutagens | Promote DNA stem-loop potential | Recombination promoted | Error detection and repair | Chargaff's second parity rule and genome-wide stem-loop potential |
| Classical phenotypic selective factors | Impair homology search between recently duplicated genes | Recombination between similar genes impaired | Gene duplication | Chargaff's \((G+C)\%\) rule |
| Differences in "reprotoype" (Recombinationally "not-self" sexual partners) | Impair homology search between DNAs of species members whose sequences are diverging | Meiotic recombination impaired | Species duplication (speciation) |  |

Table 10-1. Summary of multiple, potentially conflicting, levels of information in genomes, and their relationships to Chargaff's rules

Nucleic acid pressures can affect the success of protein alignment analyses based on the scoring schemes described in standard bioinformatics texts. For example, since the amino acid tryptophan is highly conserved in proteins but glutamic acid is not, a tryptophan match between two protein sequences scores more than a glutamic acid match between two protein sequences [28]. Thus, two protein sequences may have a low similarity score because tryptophans do not match. Where there is tryptophan in one sequence, an alignment program may score a mismatch, or place a gap, in the other sequence. However, tryptophan may be present at a certain position in a particular protein sequence because TGG (the codon for tryptophan) plays some role at the nucleic acid level that is not needed in the homologous gene from another species.
Thus, gaps in the alignment of homologous protein sequences may occur if nucleic level pressures on the corresponding DNA sequences differ. To avoid this problem, protein alignment algorithms may depend on data from closely related organisms (e.g. "PAM matrices"), or use short "blocks" of amino acids (e.g. "BLOSUM matrices") corresponding to regions that are highly conserved at the protein level (e.g. active sites of enzymes). Trade-offs between competing pressures will be further explored in Chapter 11.

**Summary**

If genome space is finite with little, if any, DNA that is not functional under some circumstance, then potential conflicts between different forms of genomic information must be resolved by appropriate trade-offs. These trade-offs include the insertion into genes of spacers, introns, and simple sequence elements. The nature and extent of the trade-offs varies with the biological species. Study of trade-offs is facilitated in genes or species where demands are extreme (e.g. genes under positive selection pressure to adapt proteins, genes that overlap, and species under extreme downward or upward GC-pressure). Spacers and introns are likely to have arisen early in evolution because they are preferential sites for the stem-loop structures in DNA that are necessary for initiating recombination and, hence, error-detection and correction. Purine-loading pressure would have supported fold pressure in provoking the splitting into introns of what might otherwise have been large exons. From this perspective we can identify the Achilles heel of the AIDS virus as the dimer-initiation sequence that is essential for the copackaging of disparate genomes, so allowing recombination repair in a future host.