HIERARCHICAL INFORMATION CONTENT, LINGUISTIC PROPERTIES AND PROTEIN-BINDING OLIGOMERS IN CODING AND NONCODING DNA SEQUENCES

DEAN H. KENYON, Ph.D.
DEPARTMENT OF BIOLOGY
SAN FRANCISCO STATE UNIVERSITY
SAN FRANCISCO, CA 94132

KEYWORDS: Coding DNA, noncoding DNA, introns, hierarchical information, redundancy, linguistic analysis, Fourier analysis, protein-binding sequences

ABSTRACT

As much as 97 percent of the DNA in mammalian genomes apparently does not code for protein amino acid sequences. Some of the noncoding DNA is known to function in various gene regulatory roles. The remainder of the noncoding DNA consists mainly of introns the functions of which are largely unknown. In this study large (72,000 base pairs) concatenated sequences of human coding and intronic DNA were analyzed by means of information theoretic and linguistic FORTRAN algorithms on a Sun Sparc 1000 system. The aim was to determine the statistical and linguistic "textures" of the two categories of DNA as a means of developing a new line of evidence that might provide a basis for an empirical distinction between an intelligent-design origin and an evolutionary origin of genomes. Calculations were run on both the natural DNA sequences and on their randomized counterparts. Similar analyses were performed on the sequenced genome of *Mycoplasma genitalium,* which consists of 88% coding DNA and does not contain introns.

The hierarchical information content of the human (concatenated) coding sequences examined in this study was 1.948-1.951 bits/nucleotide up to the dinucleotide level and 1.912-1.916 bits/nucleotide up to the pentanucleotide level. For intronic DNA the corresponding values were 1.905-1.947 and 1.876-1.901. The Shannon redundancies for the coding DNA sequences are 1.34-1.44% at the dinucleotide level and 2.70-2.83% at the pentanucleotide level. The corresponding values for intronic DNA are 1.36-3.68% and 3.04-4.84%. The linguistic vocabularies of coding and noncoding DNA sequences of comparable lengths show significant differences in preferred (standard deviate $\geq$ 3.0) oligomers and avoided (standard deviate $\leq$ -3.0) oligomers. Intronic sequences exhibit marked modulo 2 periodicities in the spacing of pairs of mirror-symmetric oligomers whereas coding sequences do not show this periodicity. Mirror-complementary oligomers are less abundant than mirror-symmetric and tandemly repeating oligomers in both the coding and noncoding DNAs. Mirror-complementary oligomers occur with higher frequencies in intronic sequences compared to their randomized counterparts than in codonic sequences compared to their randomized counterparts. Coding sequences show marked periodicities modulo 3 in the spacing of tandemly repeating oligomers, whereas the intronic sequences examined in this study do not show this periodicity. The pattern of frequencies of protein-binding sequences in introns differs from that of coding DNA.

It is concluded that significant statistical and linguistic differences exist between the coding and intronic DNA of the human genome. These results are consistent with the hypothesis that intronic DNA may play a variety of vital roles in the cell biology of development in multicellular organisms. It is plausible that these roles were present from the very beginning of the existence of organisms on the earth.

INTRODUCTION

Most of the DNA (up to 97%) in mammalian genomes apparently does not code for the amino acid sequences of proteins [20]. This noncoding DNA consists of regulatory elements near the 5'-end of genes, the 3'-untranslated regions, spacer sequences between genes, telomeric and centromeric DNA,

and introns, the intragenic segments that lie between the coding segments of genes. Some mammalian genes have no introns but the majority of them sequenced to date have one or more introns. The dystrophin gene, one of the largest in the human genome, has 78 introns [27]. Prokaryotic organisms generally lack introns although exceptions are known, especially among the archaebacteria [14]. Introns are characteristic of organisms that have complex developmental processes, i.e., the multicellular animals and plants.

Although large amounts of noncoding DNA exist in mammalian genomes, both the function and origin of this DNA remain largely unknown. Several suggestions regarding the functions of introns have been made in specific cases. These include various aspects of the regulation of gene transcription [1, 3, 29], especially in the early stages of development [16]. Evolutionists have proposed two general hypotheses regarding the origin of introns: (1) Introns are remnants of the hypothetical prebiotic "RNA world." These remnants survived into the cellular era, were somehow lost from prokaryotes but somehow retained in eukaryotes . For a critical review of the RNA-world hypothesis, see Mills and Kenyon [19]. (2) Introns developed later in evolution from codonic or some other category of DNA (the "introns-late" theory) [1, 11, 14]. A third possibility is that introns and codonic sequences were created by an intelligent designer as distinct types of DNA, each playing vital roles in the irreducibly complex machinery of the earliest eukaryotic organisms.

The direct laboratory attack on the problem of the functions of specific introns (i.e., functions besides just being excised out of the genes in which they occur) will of course continue and will provide, little by little, the information needed to develop a clearer picture of the activities of introns in general. Another approach is to treat the problem more globally by analyzing by computer large, concatenated sequences of intronic and coding DNA looking for statistical and linguistic differences in oligomer frequencies that might provide clues pertaining to the origin and function of the non-coding DNA (see e.g., Refs. [4, 13, 18, 21, 22]). If it can be shown, for example, that the statistical "texture" of noncoding DNA is significantly different from that of coding DNA, this may support the view that one did not originate from the other, but that both classes were distinct from their earliest occurrence. "Texture" as used herein refers qualitatively to the general visual appearance of a DNA sequence as it is displayed on a computer screen. For example, if a sequence contains appreciable numbers of simple repeats like $(at)_n$ or t_n or quasi-periodicities of various kinds, it has a different texture than a sequence in which no such patterns are apparent.

In this study large (~72 Kb) concatenated pools, or "modules" of human intronic DNA were assembled from GENBANK data, as well as modules of purely coding human DNA (coding exonic DNA).

This paper describes a preliminary, exploratory project employing a variety of computational techniques to search for consistent patterns of differences between codonic and intronic human DNA. The aim was to develop a sufficiently large body of statistically reliable quantitative results to allow assessment of the feasibility of an empirical test of the intelligent-design hypothesis regarding the origin of complex genetic systems.

Mycoplasma genitalium was included in this pilot study because its entire genome has been sequenced and it "...is thought to contain the smallest genome for a self-replicating organism...and represents an important system for exploring a minimal functional gene set" [10]. This organism, so far as is known, contains no intronic DNA and does not have the kind of developmental processes characteristic of multicellular organisms.

METHODS

Preparation of DNA Sequence Modules

Seven intron modules, each approximately 72,000 bases long (+ strand only), were constructed by concatenating 371 introns from 67 completely sequenced human genes obtained from GENBANK. Calculations were run on a SunSparc 1000 system.

Two coding DNA modules, each about 72,000 bases long (+ strand only) were assembled by concatenating the coding portions of 101 genes (noncoding exonic DNA was not included) either by excising and splicing the codonic sequences from complete gene files, or by using cDNA sequences (derived from mRNA).

The complete genome of *Mycoplasma genitalium* [10] was divided into 8 contiguous modules of about 72,500 bases each (+ strand only).

The sequence modules were randomized (shuffled) by means of an algorithm incorporating the ran2 function given in Press *et al.* [24]. The general features of the sequence modules are given in Table 1.

TABLE 1. DNA Sequence Modules

Module	Length*	No. of Bases*	No. of Introns or Coding Sequences**	No. of Genes**
hrb17¥	71,718	71,718	1	1
hrbintpool1	71,976	71,967	9	1
intpool2	72,212	72,162	50	5
intpool3	72,113	72,063	50	8
intpool4	72,186	72,108	78	14
intpool5	72,172	72,067	105	23
intpool6	72,070	71,992	78	21
codpool1	71,983	71,931	52	52
codpool2	72,164	72,115	49	49
mycomods† 1-6	72,509	72,509	59	59
mycomod7	72,508	72,508	59	59
mycomod8	72,511	72,511	59	595

¥Human retinoblastoma gene, intron 17. *Modules that are concatenated sequences have a spacer symbol after each intron or codonic sequence. †*Mycoplasma genitalium*. The completely sequenced genome was divided into 8 modules. **The value of 59 for mycomods 1-8 is the average number of coding sequences (= the average number of genes) per module, i.e., 470 genes in 580,073 bp [10].

Information Theoretic Analyses

The information content per base, taking into account the base composition and the frequencies of oligomers of length 2-6, was estimated using the hierarchical method of Gatlin [12, 25]. The Shannon redundancy for each sequence module, $\%R_n = 100 \times [1-H(n)/2n]$ where $H(n) = -\sum p_i \log_2 p_i$, n = oligomer length, and the summation runs from i = 1 to $i = 4^n$, was calculated for n = 1-6 [9, 18]. The larger the value of $\%R_n$ for a sequence module of length L, the greater is the degree to which the frequencies of oligomers of length n in that module differ from the frequencies expected in a randomized DNA sequence of length L containing equal numbers of the bases, a, t, c and g.

Linguistic Analyses

The methods of Beckmann *et al.* [4] were used to identify short (n = 2-6) overlapping oligomers that occur significantly more frequently or less frequently than would be expected on the basis of a random distribution of their constituent shorter oligomers. Oligomers with "standard deviates" (not to be

confused with standard deviations, which characterize distributions of numbers, e.g., the frequencies of all the 256 tetramers) ≥ 3.0 are designated "words," while those with standard deviates ≤ -3.0 are designated "antiwords." These identifications allow direct comparisons of the linguistic vocabularies of the intronic and coding DNA modules.

Global (i.e., based on deviations of observed oligomer frequencies from those expected in a random sequence containing equal numbers of each of the four bases) correlation coefficients for observed tetramer frequencies were computed for pairs of coding and intronic DNA modules using the method of Brendel *et al.* [7] (see Appendix). In this method the deviations from the expected frequencies in a random sequence (frequency/n for each base = 0.25) observed in one sequence module are compared to those observed in another sequence module.

"Local" (i.e., based on deviations of oligomer frequencies observed in a given module from the frequencies expected in the randomized counterpart of the module preserving the same base composition) correlation coefficients, C_k, where k denotes oligomer length, for pairs of sequence modules were calculated following the method of Pietrokovski *et al.* [23]. In this procedure, the standard deviates of the oligomers of given lengths are compared between two modules. From the values of C_k, k = 2-5, a similarity coefficient, S, is calculated as follows:

$$S = (C_2 + C_3 + C_4 + C_5)/4$$

Repeating Nucleotides and Dinucleotides

Intron files in the modules were scanned visually to identify the most common oligomers containing strings of repeating bases or repeating dinucleotides. The frequencies and size distributions of these (non-overlapping) oligomers were quantified by computer scanning.

Mirror-Symmetric, Mirror-Complementary, and Tandemly Repeating Oligomers

Each sequence module was scanned for the frequencies of (overlapping) pairs of mirror-symmetric, mirror-complementary, and tandemly repeating oligomers separated by distances from 0 to 22 bases. Examples of these oligomers are given below. The searches were confined to the regions between the simple repeating oligomers defined in the previous paragraph.

Mirror-symmetric: atcc----------ccta

Mirror-complementary: atcc----------ggat

Tandemly repeating: atcc----------atcc

Fourier Analysis of Tandem Periodicity

A very powerful technique for detecting certain kinds of periodicities in strings of symbols is the fast Fourier transform (FFT) given by Press *et al.* [24]. The FFT was used to estimate the power spectra (graphical representations of the relative strength of each frequency component contributing to the periodicity) for the first 65,536 (i.e., 2^{16}) bases of the coding and intronic modules using the method of Voss [28]. Special attention was given to frequencies near 0.333 (modulo 3 periodicity), and 0.500 (modulo 2 periodicity) because of reported differences between codonic and intronic DNA at these frequencies [2, 13, 17].

Consensus Sequences for Protein-Binding Sites

The sequence modules were scanned for the recognition consensus subsequences for a variety of common transcription factors [1, 21] and for other DNA-binding proteins [6]. The observed frequencies of these motifs in the codonic and intronic modules were compared with the frequencies expected in the randomized counterparts of the modules.

RESULTS

The base composition of each sequence module is given in Table 2, along with that of the entire

human genome.

TABLE 2. Base Composition of Sequence Modules*

Module	a	t	c	g
hrb17	0.3136	0.3199	0.1786	0.1879
hrbintpool1	0.2933	0.3229	0.1899	0.1941
intpool2	0.2822	0.2924	0.2078	0.2176
intpool3	0.2561	0.2770	0.2255	0.2414
intpool4	0.2256	0.2363	0.2679	0.2702
intpool5	0.2336	0.2535	0.2454	0.2673
intpool6	0.2863	0.3003	0.1999	0.2135
intmods combined	0.2701	0.2860	0.2165	0.2275
human genome¥	0.303	0.303	0.199	0.195
codpool1	0.2643	0.2157	0.2553	0.2647
codpool2	0.2551	0.2152	0.2690	0.2608
mycomods combined	0.3457	0.3374	0.1579	0.1591

*(+) strand only, except for the human genome. ¥Liver [8].

Hierarchical Information Content and Redundancy Profiles

The results of hierarchical information theoretic analyses of the sequence modules are given in Table 3. These data show that the redundancy in the intronic modules is higher than in the coding modules, and correspondingly, the hierarchical information content of the intronic sequences is lower than that of the coding sequences. The difference in redundancy is due in part to differences in base composition between intronic (a + t = 0.56) and codonic (a + t = 0.48) DNA. The redundancy in the mycoplasmal sequences is considerably higher than in both the codonic and the intronic DNA modules as expected from the high a + t content (0.68) of the mycoplasmal DNA. The genome of *Mycoplasma genitalium* contains approximately 88% coding and 12% noncoding (but not intronic) DNA [10].

Examples of the redundancy profiles (cf. Mantegna *et al.* [18]) are shown in Figures 1 and 2. Figure 1 shows the redundancy profile for intron module 4 (78 introns in 14 genes) with its randomized counterpart, together with that of codon module 2, and its randomized counterpart. It is apparent that the redundancy of intron module 4 is significantly greater than that of codon module 2. A similar result was obtained when codon module 1 was substituted for codon module 2. In Figure 2 the curves for intpool5 and codpool1 are translated along the vertical axis to coincide with the curves for intpool6 at n = 1. In this way the redundancy differences due solely to the differences in base composition among the modules is eliminated and the "excess redundancy" for oligomer lengths 2-6 can be easily visualized. Figure 2 shows that intron modules 5 (105 introns in 23 genes) and 6 (78 introns in 21 genes) have greater excess redundancies than codon module 1. Substitution of codon module 2 for codon module 1 gave a similar result.

TABLE 3. Hierarchical Information Content and Shannon Redundancies of Sequence Modules

Modules	HM2* (bits/base)	HM5* (bits/base)	%R2	%R5
Human intronic	1.905-1.947	1.876-1.901	1.36-3.68	3.04-4.84
Human Coding	1.948-1.951	1.912-1.916	1.34-1.44	2.70-2.83
Mycomods 7,8	1.836-1.837	1.804-1.805	7.38-7.39	8.47-8.52

*HM2 = information content per base considering base composition and dinucleotide frequencies. HM5 = information content considering base composition and the frequencies of oligomers of length 2-5 [12].

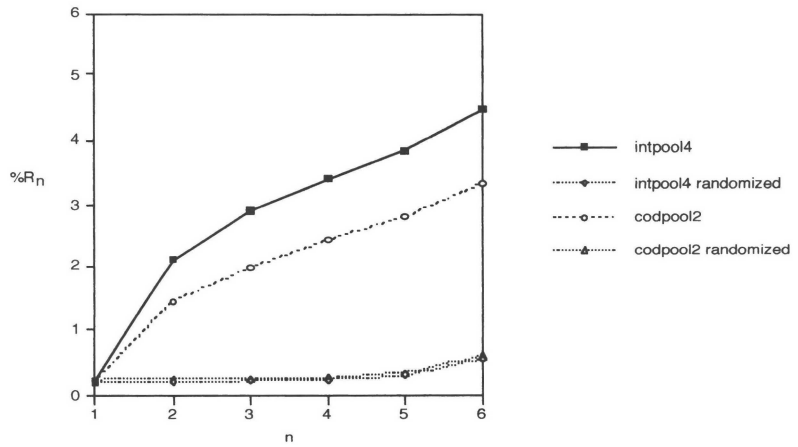

Fig. 1. $\%R_n$ v. n for intron module intpool4 and codon module 2 and their randomized counterparts.

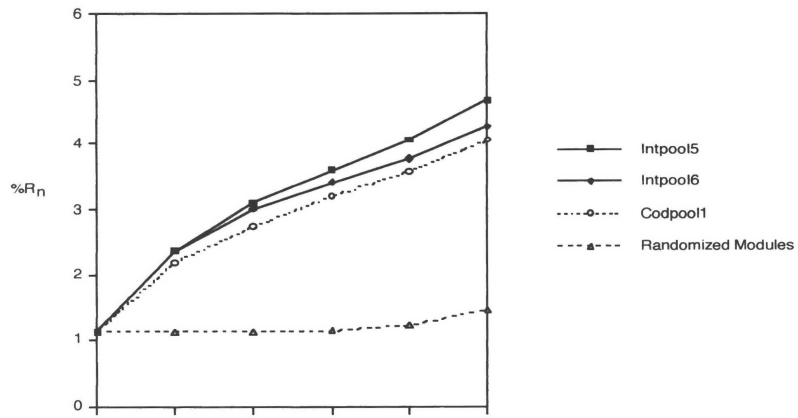

Fig. 2. $\%R_n$ v. n for intron modules 5 and 6 and codon module 1 and their randomized counterparts.

TABLE 4. The Most Preferred and Most Avoided Dinucleotides and Trinucleotides in the Sequence Modules*

	hrb17	hrbintpool1	intpool2	intpool3	intpool4	intpool5	intpool6	codpool1	codpool2	mm7¥	mm8¥
	tt	cc	cc	cc	ct	gg	cc	tg	tg	aa	aa
	cc	gg	gg	gg	ag	cc	gg	ct	ct	tt	tt
Preferred	ca	tt	ag	ag	tg	ct	ag	ca	ca	gc	gc
	gg	aa	tt	tg	cc	ca	ct	ag	cc	ca	ca
	ag	ca	ct	ca	gg	ag	ca	ga	ag	tg	tg
	--	--	--	--	--	--	--	--	--	--	--
	--	--	--	--	--	--	--	--	--	--	--
	ac	at	at	at	gt	at	at	at	gc	ga	ga
	gt	ac	ac	gt	at	ac	gt	ac	ac	ac	ac
Avoided	at	gt	gt	ac	ac	gt	ac	gt	gt	cg	at
	ta	ta	ta	ta	ta	ta	ta	ta	ta	at	ta
	cg	cg	cg	cg	cg	cg	cg	cg	cg	ta	cg
	ttt	aaa	aaa	aaa	aaa	ttt	ttt	tac	tac	atc	atc
	aaa	ttt	ttt	ttt	ttt	aaa	cag	tat	tat	acc	gat
Preferred	ata	ata	ctc	ata	gtg	gtg	aaa	ttc	aag	gtt	acc
	tat	tat	cac	ctg	aca	cac	gtg	tgg	ttc	gat	gtt
	ctg	cag	gag	ctc	gag	ata	ata	ctg	tgg	tca	taa
	--	--	--	--	--	--	--	--	--	--	--
	--	--	--	--	--	--	--	--	--	--	--
	cta	ctt	gtt	cta	tcg	tag	tag	cga	cga	gcc	ctc
	gat	cta	ctt	ctt	cta	gtt	cta	ggg	ggg	ctc	gac
Avoided	aag	aag	ttg	aag	aag	cta	gtt	ccc	aat	ggc	ggc
	ttg	ttg	aag	caa	ttg	caa	ttg	ttg	ttg	aca	gtc
	caa	caa	caa	ttg	caa	ttg	caa	tag	tag	gtc	aca

*The five most preferred and most avoided oligomers are listed in the order of descending values of the standard deviate. ¥mm7 and mm8 denote mycoplasmal modules 7 and 8.

Linguistic Analyses

Table 4 lists the five most preferred and the five most avoided dinucleotides and trinucleotides in the intron and codon modules in the order of descending values of the standard deviate. The corresponding lists for mycoplasmal modules 7 and 8 are included for comparison. At the dinucleotide level the human codonic and mycoplasmal antiword lists closely resemble those of the intronic modules. For example, ta and cg are the two most avoided dinucleotides in all the modules. The codonic preferred words show an intermediate degree of similarity to the intronic preferred words, while the mycoplasmal preferred words are less similar to the intronic words. At the trinucleotide level there is more resemblance between the codonic and intronic antiword lists than between the codonic and intronic preferred word lists. The mycoplasmal preferred word lists have no overlap with the codonic and intronic lists. The oligomers ttt and aaa are the two most preferred trinucleotides in all the intron modules but are absent from the codonic and mycoplasmal lists.

The average values of the global correlation coefficients, r (for tetranucleotides), and similarity coefficients, S, for intron/intron, intron/coding, intron/mycoplasma, and coding/mycoplasma pairs of modules are given in Table 5. Both coefficients have much higher values for the intron-intron comparisions than for the intron-coding comparisons. The global correlation coefficient for the intron/mycoplasma pairs of modules is higher than the value for the intron/coding pairs. The oligomer vocabularies of the coding modules show very little correlation with mycoplasmal vocabularies by either measure.

TABLE 5. Average Values of Correlation and Similarity Coefficients for the Oligomer Frequencies in Pairs of Sequence Modules

Module Pairs	Global Correlation Coefficient, r	Similarity Coefficient, S
intron/intron	0.74 (21)*	0.76 (21)*
intron/coding	0.48 (14)	0.46 (14)
intron/mycoplasma	0.55 (7)	0.07 (14)
coding/mycoplasma	-0.04 (2)	0.12 (2)

* Numbers in parentheses indicate the number of values averaged in each case.

Motifs in Intronic and Coding DNA

Direct inspection of the intronic DNA sequences revealed the presence of many repeating oligomers of the form a_n, t_n, c_n, g_n, $(at)_n$, $(ag)_n$, $(ca)_n$, $(ct)_n$, and $(gt)_n$, $n \geq 6$ for the mononucleotides and $n \geq 3$ for the dinucleotides. Such oligomers occur only rarely, if at all, in the codonic modules. The presence of these highly ordered oligomers, as well as more complex periodic and quasi-periodic sequences (e.g., the short interspersed elements, SINES, such as the Alu family, and the long interspersed elements, LINES) have long been noted in noncoding DNA [15, 20].

Figure 3 shows the size distribution of non-overlapping t_n oligomers in intronic and coding DNA. A similar distribution was observed for the a_n oligomers. The frequencies of the c_n and g_n oligomers and of the repeating dinucleotides were all much lower than the frequencies of the a_n and t_n oligomers. Table 6 summarizes these data for the three classes of modules. The a_n and t_n oligomers and the repeating dinucleotides are much more abundant in the intronic modules than in the codonic modules. The a_n and t_n motifs, especially for n = 6-8, are more abundant in the mycoplasmal modules than in the intronic and coding modules, whereas the c_n, g_n, and all of the repeating dinucleotide motifs are absent from or very rare in the mycoplasmal DNA. These results are perhaps not surprising given the high a + t content (0.68) of the mycoplasmal modules.

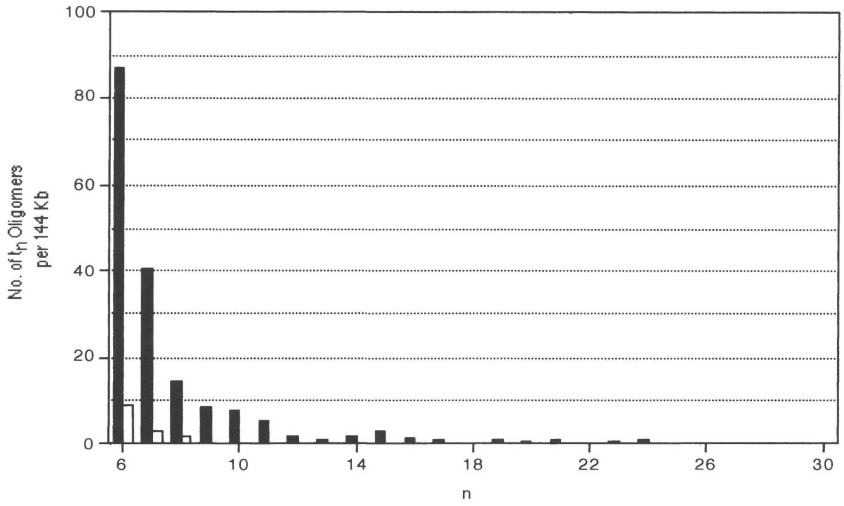

Fig. 3. Size distribution of t_n oligomers in intronic and codonic DNA.

Intronic DNA ■ □ Coding DNA

The frequencies of mirror-complementary heptamers spaced a distance d apart are shown in Figure 4 for the two codon modules combined and for the seven intron modules normalized to 144 kilobases. There is no clear pattern of periodicity of spacing of these oligomers in the coding DNA modules. The intronic DNA shows a weak modulo 2 periodicity for d values from 0 to 11. Mirror-complementary heptamers are about two-fold more abundant in the intronic DNA than in the coding DNA.

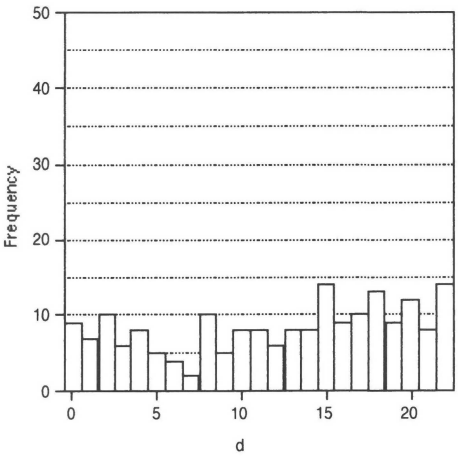

Fig. 4A. Frequency of pairs of mirror-complementary heptamers spaced d bases apart in 144 Kb of coding DNA.

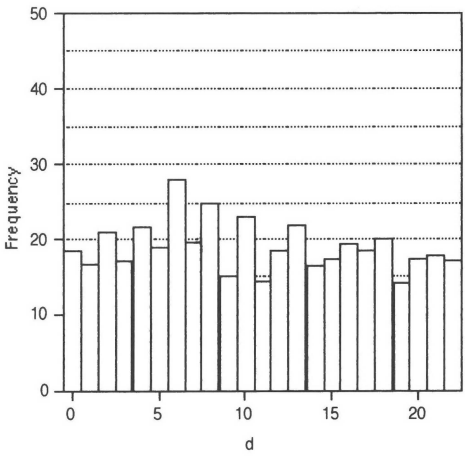

Fig. 4B. Frequency of pairs of mirror-complementary heptamers spaced d bases apart in the combined intron modules. Data normalized to 144 Kb.

Some of these oligomer pairings as they would occur in the corresponding RNA transcripts, e.g., those with d values from 6 to 20, may have the

potential of forming loop-stem structures.

Figure 5 shows the frequencies of mirror-symmetric heptamers spaced a distance d apart in coding and intronic DNA. Such oligomers are far more abundant in introns than in coding DNA. While there is no clear pattern of periodicity in the data for mirror-symmetric heptamers in the coding DNA, the intronic DNA shows a clear modulo 2 periodicity which declines in strength with increasing values of d.

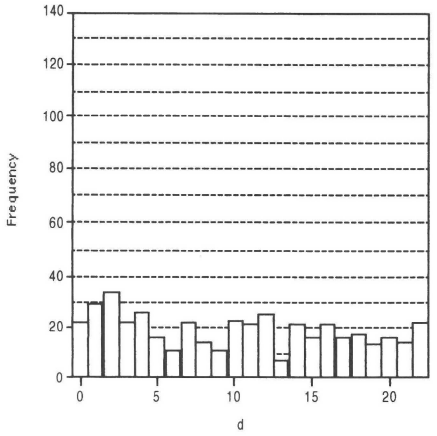

Fig. 5A. Frequency of pairs of mirror-symmetric heptamers spaced d bases apart in 144 Kb of coding DNA.

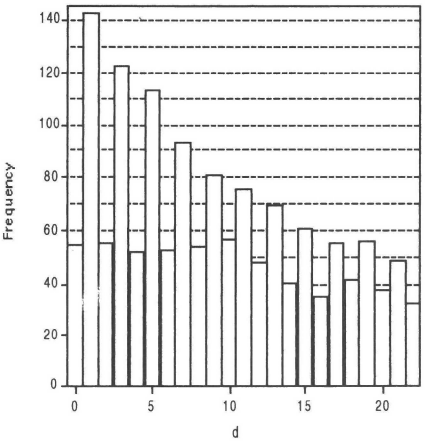

Fig. 5B. Frequency of pairs of mirror-symmetric heptamers spaced d bases apart in the combined intron modules. Data normalized to 144 Kb.

The frequencies of tandemly-repeating heptamers spaced a distance d apart are shown in Figure 6 for coding and intronic DNA. While the overall abundance of such oligomers is higher in the intronic than in the codonic DNA, the latter exhibits a clear modulo 3 periodicity of spacing. No such periodicity is seen in the intronic DNA.

The strong modulo 3 periodicity seen in the spacing of tandem repeats in coding DNA is corroborated by the approximate Fourier power spectra obtained by the fast Fourier transform (FFT). Figure 7 shows the power spectrum for the first 65,536 (2^{16}) bases of codon module 1 and the corresponding spectrum for intron module hrb17. The spectrum for the coding DNA has a single prominent peak at j = 21,845 corresponding to a period of 3.00005 bases. The peak is absent from the intron spectrum. Neither type of DNA exhibited a peak in the region of period 2.0.

Recognition Sequences for Transcription Factors and Other DNA-Binding Proteins

The pattern of occurrence of recognition sequences for DNA-binding proteins in the coding and intronic modules is shown in Table 7. The number of occurrences of each recognition sequence in the DNA modules is compared to the number of occurrences expected in the randomized (shuffled) counterparts of the modules. The data demonstrate that there are significant differences in the patterns of occurrence of the consensus oligomers between coding and intronic DNA. For example, gtataa is more strongly avoided in coding DNA than in intronic DNA, whereas the pattern is reversed in the case of tgacgt. The frequency of ataaag in the intronic modules is about the same as would be expected in the randomized modules, whereas the frequency of this recognition sequence in the coding modules is higher than would be expected in the randomized codonic modules.

DISCUSSION AND CONCLUSIONS

The human intronic DNA sequences are more redundant, more highly ordered, and of lower information content (measured in bits/base) than the human coding sequences (Table 3, Figures 1 and 2). The higher degree of order in introns may in part be explained by the presence of simple repeating patterns of nucleotides such as a_n, t_n, $(at)_n$, and $(ca)_n$ (Figure 3, Table 6). While such repeating motifs are rare in codonic DNA since they would clearly interfere with the protein-coding function of this DNA, they may well have intron-specific functions yet to be identified. On this hypothesis they might tentatively be termed "intron tools." In any case, such oligomers are among the most striking

characteristic features of intronic DNA.

As demonstrated by both the global correlation and similarity coefficients given in Table 5, the linguistic vocabulary of intronic DNA differs markedly from that of codonic DNA. Moreover, the frequencies of mirror-symmetric pairs of oligomers are much higher in intronic than in coding DNA (Figures 5A and 5B), and these oligomers display a strong modulo 2 periodicity of spacing in the stretches of intronic DNA between the "tool" oligomers (a_n, t_n, $(at)_n$, etc.). Such a periodicity does not occur in the coding modules. It is conceivable that the abundant mirror-symmetric oligomers in introns might play some role in intron processing related to the speed and accuracy of splicing and subsequent degradation of the excised intronic sequences (see e.g., Beckmann *et al.* [4, p.399]).

TABLE 6. Total and Average Lengths of Non-Overlapping Simple Repeating Oligomers in Sequence Modules

	Intronic Modules			Coding Modules			Mycoplasmal Modules		
Oligomer	TL*	L¥	%‡	TL	L	%	TL	L	%
a_n	1297	7.9	3.3	353	6.2	0.9	2526	6.4	5.0
t_n	1385	7.8	3.4	91	6.5	0.3	3101	6.5	5.9
c_n	122	6.3	0.4	271	6.6	0.7	14	7.0	0.06
g_n	163	6.2	0.5	116	6.1	0.3	0	0	0
$(at)_n$	120	9.1	a: 0.15 t: 0.15	0	0	0	0	0	0
$(ag)_n$	96	9.9	a: 0.12 g: 0.15	60	8.6	a: 0.08 g: 0.08	8	8.0	a: 0.01 g: 0.02
$(ca)_n$	129	13.2	c: 0.21 a: 0.17	32	8.0	c: 0.04 a: 0.04	0	0	0
$(ct)_n$	77	9.3	c: 0.12 t: 0.09	16	8.0	c: 0.02 t: 0.03	0 0	0 0	0 0
$(gt)_n$	137	11.2	g: 0.21 t: 0.17	8	8.0	g: 0.01 t: 0.01	8	8.0	g: 0.02 t: 0.01
∑ TL**	3526			947			5657		
T%¥¥	2.45			0.66			3.9		

*TL = total length of the oligomers of a given type in the modules. The values for the intron modules are normalized to 144 Kb for direct comparison with the other types of modules. ¥L = average length of the oligomers of a given type in the modules. ‡% = percentage of the total number of bases of a given type (a, t, c, g) that occur in the given oligomer in the modules (e.g., 3.4%, or 4847, of the total number of thymines in the combined intron modules, 144146, occur in non-overlapping oligomers, t_n, where $n \geq 6$). **∑ TL = total number of bases that occur in the listed simple repeating oligomers in the modules. ¥¥ T% = percentage of the bases in the combined modules of a given type that occur in the listed oligomers.

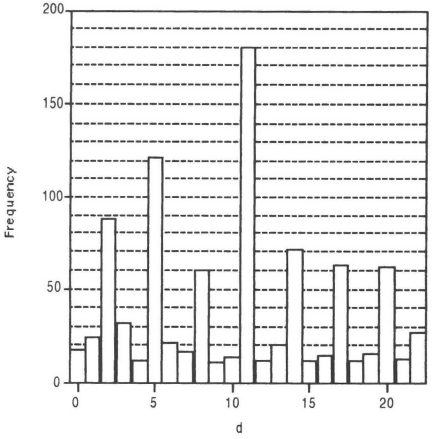

Fig. 6A. Frequency of tandemly-repeating heptamers spaced d bases apart in 144 Kb of coding DNA.

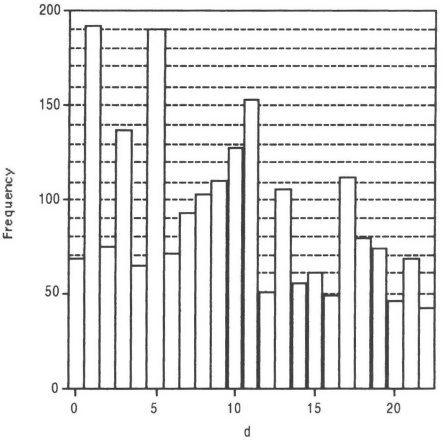

Fig. 6B. Frequency of tandemly-repeating heptamers spaced d bases apart in intronic DNA. Data normalized to 144 Kb.

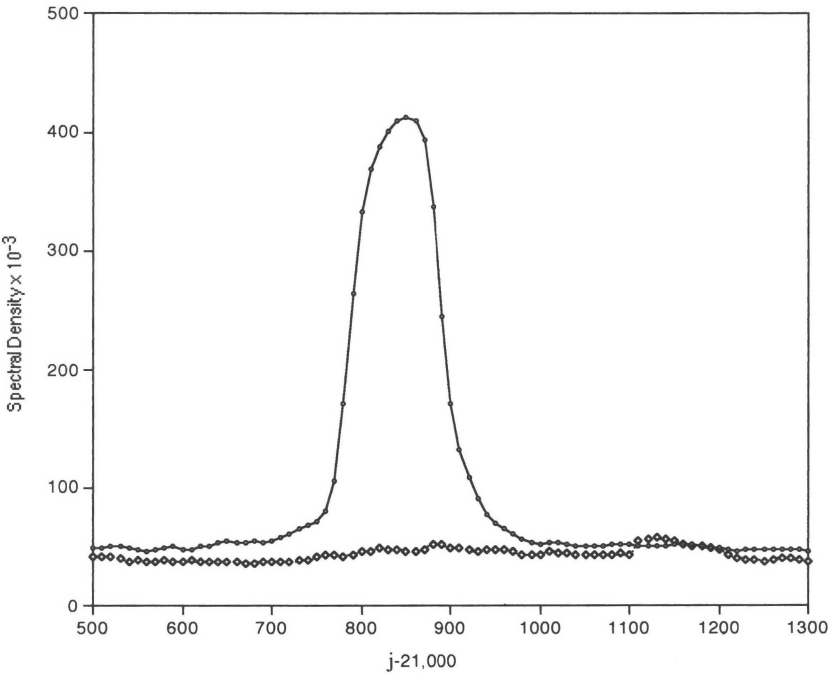

Fig. 7. Fourier power spectra for the first 65,536 (nb) bases of codonic module 1 and intronic module hrb17. j = frequency x nb. Period = nb/j. The coding DNA peak occurs at j = 21,845 (period = 3.00005 bases). The data were smoothed by averaging the spectral density values from j - 50 to j + 50 for each value of j.

Codon Module 1 ..... Intron Module hrb17

TABLE 7. Frequencies of Protein-Binding Consensus Sequences in Coding and Intronic DNA

Consensus Sequence	Protein*	Ref.	Coding Modules**			Intronic Modules***		
			Freq.	LR‡	Freq./LR	Freq.	LR	Freq./LR
tcatgag	Hepta/Rev	[1]	14	8.2	1.71	11.1	9.1	1.22
ataaag	GATA-1	[21]	49	37.1	1.32	52.9	51.9	1.02
ccgccc	BGP1_RS/Rev	[1]	49	47.0	1.04	18.0	20.2	0.89
aaagtgt	CAP/CRP-lac	[1]	8	8.1	0.99	14.0	11.8	1.19
topoII¥	Topoisom. II	[26]	6	7.2	0.83	5.7	7.6	0.75
tataaa	TFIID	[21]	24	30.4	0.79	84.0	73.4	1.14
gattgg	CTF/CBP-hs	[1]	22	31.4	0.70	26.6	35.9	0.74
ctataa	TFIID	[21]	19	30.6	0.62	40.0	52.2	0.77
tgacgt	ATF_RS	[1]	16	31.3	0.51	6.9	34.2	0.20
ttatat	GAL1-TATA/Rev	[1]	10	20.8	0.48	70.9	81.6	0.87
acgtca	CREB	[21]	17	37.6	0.45	7.1	30.5	0.23
gtataa	TFIID	[21]	10	30.7	0.33	33.7	54.9	0.61

*All proteins listed are transcription control factors except topoisomerase which modifies the supercoiling of DNA. **Codpools1 and 2. *** Data from the seven intron modules normalized to 144 Kb. ‡LR = frequency expected in the randomized modules. ¥Topoisomerase II = rnynncnngyngktnnyny, where r = a or g, n = any of the four bases, y = c or t, and k = g or t.

Coding DNA exhibits a marked modulo 3 periodicity in the spacing of pairs of tandem repeats (Figure 6A), a pattern not seen in the intronic DNA (Figure 6B). The Fourier power spectra (Figure 7) further emphasize this sharp distinction between the two categories of DNA. The fast Fourier transform detects tandem periodicities that are "densely packed" and ubiquitous throughout the concatenated DNA sequences. It does not detect the kind of modulo 2 periodicity in mirror-symmetric oligomers found in intronic DNA. In general the FFT does not detect local tandem periodicities separated by long streches of aperiodic sequence. These are lost in the noise of the spectra.

If introns evolved from coding DNA, then it is difficult to see why all traces of the modulo 3 periodicity of the parent material should now be absent from the intronic sequences. If, on the contrary, both codonic and intronic DNA were present in multicellular organisms from the beginning of their existence on the earth, then the current absence of modulo 3 periodicity in intronic DNA would be explained. It is of course conceivable that introns may have developed from codonic DNA originally present in the intronless genomes of primitive mycoplasma-like organisms and then lost all traces of the codonic tandem periodicity by subsequent evolutionary events. Nevertheless, when all the differences between these two types of DNA are considered, it is at least as likely that intronic and codonic DNA are two functionally distinct, necessary components of the irreducibly complex [5] molecular biological apparatus of eukaryotic cells.

The fact that the recognition consensus sequences for many transcription factors and other proteins that interact with DNA occur in introns with frequency patterns differing from those in the coding sequences suggests that introns may indeed play unique roles in the timing, rate and accuracy of gene expression.

ACKNOWLEDGMENT

The author thanks the Discovery Institute, Seattle, Washington for its generous support of this work in the form of an unrestricted Research Fellowship.

REFERENCES

[1] Ala-Kokko, L., Kvist, A. P., Metsaranta, M., Kivirikko, K. I., de Crombrugghe, B., Prockop, D. J. and Vuorio, E., **Conservation of the Sizes of 53 Introns and Over 100 Intronic Sequences for the Binding of Common Transcription Factors in the Human and Mouse Genes for Type II Procollagen (COLA2),** Biochemical Journal, 308:3 (1995) pp. 923-929.

[2] Arquès, D. G. and Michel, C. J., **Periodicities in Coding and Noncoding Regions of the Genes**, Journal of Theoretical Biology, 143 (1990) pp. 307-318.

[3] Baldi, P., Brunak, S., Chauvin, Y. and Krogh, A., **Naturally Occurring Nucleosome Positioning Signals in Human Exons and Introns**, Journal of Molecular Biology, 263:4 (1996) pp. 503-510.

[4] Beckmann, J. S., Brendel, V. and Trifonov, E. N., **Intervening Sequences Exhibit Distinct Vocabulary**, Journal of Biomolecular Structure and Dynamics, 4:3 (1986) pp. 391-400.

[5] Behe, M., Darwin's Black Box: The Biochemical Challenge to Evolution, 1996, The Free Press, New York, NY.

[6] Bodnar, J. W. and Ward, D. C., **Highly Recurring Sequence Elements Identified in Eukaryotic DNAs by Computer Analysis are Often Homologous to Regulatory Sequences or Protein Binding Sites**, Nucleic Acids Research, 15:4 (1987) pp. 1835-1851.

[7] Brendel, V., Beckmann, J. S. and Trifonov, E. N., **Linguistics of Nucleotide Sequences: Morphology and Comparison of Vocabularies**, Journal of Biomolecular Structure and Dynamics 4:1 (1986) pp. 11-21.

[8] Chargaff, E., **Isolation and Composition of the Deoxypentose Nucleic Acids and of the Corresponding Nucleoproteins**, The Nucleic Acids: Chemistry and Biology, E. Chargaff and J.N. Davidson, Editors, 1955, Academic Press, New York, NY, Volume 1, pp. 307-371.

[9] Chatzidimitriou-Dreismann, C. A., Streffer, R. M. F. and Larhammar, D., **Lack of Biological Significance in the 'Linguistic Features' of Noncoding DNA –A Quantitative Analysis**, Nucleic Acids Research, 24:9 (1996) pp. 1676-1681.

[10] Fraser, C. M. *et al.,* **The Minimal Gene Complement of *Mycoplasma genitalium,*** Science, 270 (1995) pp. 397-403.

[11] Gamulin, V., Skorokhod, A., Kavsan, V., Muller, I. M. and Muller, W. E., **Experimental Indication in Favor of the Introns-Late Theory: the Receptor Tyrosine Kinase Gene from the Sponge *Geodia cydonium***, Journal of Molecular Evolution, 44:3 (1997) pp. 242-252.

[12] Gatlin, L. L., Information Theory and the Living System, 1972, Columbia University Press, New York, NY.

[13] Konopka, A. K., Smythers, G. W., Owens, J. and Maizel, J. V. Jr., **Distance Analysis Helps to Establish Characteristic Motifs in Intron Sequences**, Gene Analysis Techniques, 4 (1987) pp. 63-74.

[14] Lambowitz, A. M. and Belfort, M., **Introns as Mobile Genetic Elements**, Annual Reviews of Biochemistry, 62 (1993) pp. 587-622.

[15] McConkey, E. H., Human Genetics: The Molecular Revolution, 1993, Jones and Bartlett, Boston.

[16] McKenzie, R. W. and Brennan, M. D., **The Two Small Introns of *Drosophila affinidisjuncta* Adh Gene are Required for Normal Transcription**, Nucleic Acids Research, 24:18 (1996) pp. 3635-3642.

[17] Makeev, V. I., Frank, G. K. and Tumanian, V. G., **Statistics of Periodic Regularities in Sequences of Human Introns**, Biofizika, 41:1 (1996) pp. 241-246.

[18] Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H. E., **Linguistic Features of Noncoding DNA Sequences**, Physical Review Letters, 73:23 (1994) pp. 3169-3172.

[19] Mills, G. and Kenyon, D., **The RNA World: A Critique**, Origins and Design, 17:1 (1996) pp. 9-14.

[20] Nowak, R., **Research News: Mining Treasures from 'Junk DNA,'** Science, 263 (1994) pp. 608-610.

[21] Pesole, G., Attimonelli, M. and Saccone, C., **Linguistic Approaches to the Analysis of Sequence Information**, Trends in Biotechnology, 12 (1994) pp. 401-408.

[22] Pietrokovski, S., **Comparing Nucleotide and Protein Sequences by Linguistic Methods**, Journal of Biotechnology, 35 (1994) pp. 257-272.

[23] Pietrokovski, S., Hirshon, J. and Trifonov, E. N., **Linguistic Measure of Taxonomic and Functional Relatedness of Nucleotide Sequences**, Journal of Biomolecular Structure and Dynamics, 7:6 (1990) pp. 1251-1268.

[24] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P., Numerical Recipes in FORTRAN: The Art of Scientific Computing, Second Ed., 1992, Cambridge University Press, Cambridge, pp. 498-502.

[25] Sibbald, P. R. , Banerjee, S. and Maze, J., Letter to the Editor: **Calculating Higher Order DNA Sequence Information Measures**, Journal of Theoretical Biology, 136 (1989) pp. 475-483.

[26] Spitzner, J. R. And Muller, M. T., **A Consensus Sequence for Cleavage by Vertebrate DNA Topoisomerase II**, Nucleic Acids Research, 16:12 (1988) pp. 5533-5556.

[27] Tennyson, C. N., Klamut, H. J. and Worton, R. G., **The Human Dystrophin Gene Requires 16 Hours to be Transcribed and is Cotranscriptionally Spliced**, Nature Genetics, 9 (1995) pp. 184-190.

[28] Voss, R. F., **Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences**, Physical Review Letters, 68:25 (1992) pp. 3805-3808.

[29] Wetterberg, I, Bauren, G. and Wieslander, L., **The Intranuclear Site of Excision of Each Intron in Balbiani Ring 3 pre-mRNA is Influenced by the Time Remaining to Transcription Termination and Different Excision Efficiencies for the Various Introns**, RNA, 2:7 (1996) pp. 641-651.

APPENDIX

The standard deviate of a given oligomer,s, is given by the following equation:

$$std(s) = \{f(s) - E(s)/max\sqrt{E(s),1}\}$$

where:

f(s)	=	observed frequency of oligomer, s
E(s)	=	expected frequency of oligomer, s, based on the theory of Markov processes [22]
max	=	the FORTRAN max function (i.e., the denominator is the greater of the terms, E(s) and 1)

The global correlation coefficient, r, is defined as follows:

$$r = \frac{\sum[f_1(N_iN_jN_kN_l) - n_1/256][f_2(N_iN_jN_kN_l) - n_2/256]}{\{\sum[f_1(N_iN_jN_kN_l) - n_1/256]^2 \sum[f_2(N_iN_jN_kN_l) - n_2/256]^2\}^{1/2}}$$

where:

$f_1(N_iN_jN_kN_l)$ = observed frequency of oligomer ($N_iN_jN_kN_l$) in sequence 1

n_1 = number of bases in sequence 1

$f_2(N_iN_jN_kN_l)$ = observed frequency of oligomer ($N_iN_jN_kN_l$) in sequence 2

n_2 = number of bases in sequence 2

The summation runs from 1 to 256.