Statistical features of human exons and their flanking regions

M. Q. Zhang

Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, USA

Received 24 November, 1997; Revised and Accepted 15 February, 1998

To facilitate gene finding and for the investigation of human molecular genetics on a genome scale, we present a comprehensive survey on various statistical features of human exons. We first show that human exons with flanking genomic DNA sequences can be classified into 12 mutually exclusive categories. This classification could serve as a standard for future studies so that direct comparisons of results can be made. A database for eight categories (related to human genes in which coding regions are split by introns) was built from GenBank release 87.0 and analyzed by a number of methods to characterize statistical features of these sequences that may serve as controls or regulatory signals for gene expression. The statistical information compiled includes profiles of signals for transcription, splicing and translation, various compositional statistics and size distributions. Further analyses reveal novel correlations and constraints among different splicing features across an internal exon that are consistent with the Exon Definition model. This information is fundamental for a quantitative view of human gene organization, and should be invaluable for individual scientists to design human molecular genetics experiments.

INTRODUCTION

Almost all the nuclear genes coding for proteins in eukaryotes are split into exon and intron sequences. Thus questions such as ‘what makes an exon an exon?’ and ‘how is an exon recognized by the gene expression machinery?’ are of major importance to the understanding of gene expression and regulation. The task of delineating exon–intron organization is even more challenging in vertebrates than in lower eukaryotes because an average vertebrate gene consists of multiple small exons separated by introns that are 10 or 100 times larger. As the Human Genome Project enters a large-scale sequencing phase, identifying exons has also become a bottle-neck in genome annotation. In the early 1980s, the splicing site consensus (1) and the weight matrix method (2) were developed by DNA sequence comparisons. Senapathy *et al.* (3) later compiled more comprehensive sequence statistics on major categories of GenBank release 57.0. The statistical features of promoters (4) and exon/intron size distributions (5,6) have also been studied carefully for vertebrates. There have been many good reviews on important aspects of gene recognition methods (7–9) and on assessment of different protein-coding measures (10).

To take advantage of a much larger set of human-specific sequence data available today, to facilitate experiments on human molecular genetics and to meet the need for developing better human exon recognition methods, we have extended our statistical analysis of fission yeast genes (11) to human exons and their flanking regions. Recently, accumulating experimental evidence has led to the Exon Definition model (12), which argues that, in vertebrate pre-mRNAs with large introns, the initial recognition unit of splicing is an exon defined by the interactions of splicing factors across the exon. This implies that splicing signals may be correlated across an exon and they cannot be recognized as independent sequence features. In this survey, we report our results on systematic analyses of many individual features, and we demonstrate the existence of some novel correlations and constraints among different features that may be relevant to human exon recognition and to understanding of gene expression and regulation. As the central theme in molecular biology is the structure–function relationship, distinct statistical sequence structures can often suggest, or ought to be explained by, their functions. Putting various gene sequence information in one place will help to speed up this functional interpretation.

EXON CLASSIFICATION

Exons are classified into the following 12 categories (Fig. 1), according mainly to what transcriptional or translational boundaries an exon contains [we shall refer to the poly(A) site as the end of the last exon]: (1) a 5uexon is the 5′-terminal untranslated exon in a gene; (2) a 3uexon is the 3′-terminal untranslated exon; (3) a 5utexon is the 5′-terminal exon having a 5′-untranslated region (5′UTR) followed by a coding sequence (CDS); (4) a 3tuexon is the 3′-terminal exon having a 3′UTR following a CDS; (5) an iutexon is an internal exon having a 3′ portion of the 5′UTR followed by a CDS; (6) an ituexon is an internal exon having a 5′ portion of 3′UTR following a CDS; (7) an iuexon is an internal untranslated exon; (8) an itexon is an internal translated exon. An exon in categories 9 and 10 has to contain the complete CDS: (9)

Tel: +1 516 367 8393; Fax: +1 367 8461; Email: mzhang@cshl.org

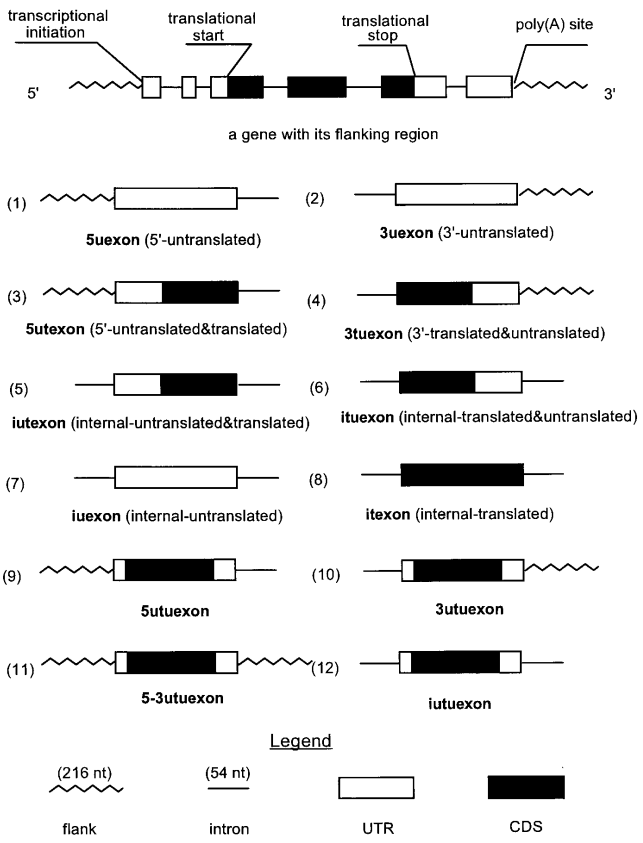
transcriptional initiation
translational start
translational stop
poly(A) site
5'
3'
a gene with its flanking region
(1)
5uexon (5'-untranslated)
(2)
3uexon (3'-untranslated)
(3)
5utexon (5'-untranslated&translated)
(4)
3tuexon (3'-translated&untranslated)
(5)
iutexon (internal-untranslated&translated)
(6)
ituexon (internal-translated&untranslated)
(7)
iuexon (internal-untranslated)
(8)
itexon (internal-translated)
(9)
5utuexon
(10)
3utuexon
(11)
5-3utuexon
(12)
iutuexon
Legend
(216 nt)
flank
(54 nt)
intron
UTR
CDS

Figure 1. Exon classification. All exons can be classified into these 12 mutually exclusive classes. On the top, a schematic gene model is depicted which indicates how some types of exons may be organized in a gene.

a 5utuexon does not contain the transcriptional end; (10) a 3utuexon does not contain the transcriptional start; (11) a 5–3utuexon contains both; and (12) an iutuexon contains neither. Because of annotational ambiguities in distinguishing between truly intronless CDSs and mRNAs, all the analyses reported in this study were done for the first eight categories, which consist of 271 5uexons, 38 3uexons, 482 5utexons, 553 3tuexons, 174 iutexons, 69 ituexons, 34 iuexons and 3440 itexons. Up until now, the focus of study has been mainly itexons, for obvious reasons. As the human genome will be completely sequenced, it is time to address issues related to all types of exons. This is the first systematic classification of exons which could serve as a standard for future studies so that direct comparison of results can be made.

STATISTICAL CHARACTERIZATION OF INDIVIDUAL SEQUENCE FEATURES

Size and compositional characteristics

Exon size distributions. The size distributions of exons that have a definite size (no '>'or '<' in their annotation) in the different categories as well as the corresponding quantile statistics are plotted in Figure 2. These results indicate that, in general, 5utexons or iuexons are relatively short (mostly <100 nt) and 3tuexons and 3uexons are relatively long (mostly –300–500 nt). While itexon sizes have a tight log-normal distribution centered around log10 (130 nt), the sizes of 3uexons are extremely heterogeneous (all >100 nt). There also seems to be a sharp drop-off for iutexon sizes >200 nt and for 3uexon sizes <100 nt. The extreme values, although somewhat dependent upon the data set, do give some idea about possible size constraints. There seems to be no minimum constraint on the size of an itexon: we found the smallest (4 nt) was the exon 3 of the human *TNNI1* gene in the initial human exon data set, although, after the data cleaning process, the minimum in our final itexon collection was 15 nt. These distributions are very useful, for example in comparative studies or in exon-trapping experimental designs.

Coding fraction and UTR distributions. The coding fraction is defined as the ratio of the CDS size over the exon size. The histograms of the coding fractions of the exons in four categories are plotted in Figure 3. These data clearly suggest that (i) translation is equally likely to start anywhere in the first exon, but it is more likely to start near the beginning of an internal exon; (ii) translation is more likely to stop near the beginning of the last exon, but it is more likely to stop near the end of an internal exon. At this point, one could only speculate on the biological implications. For example, why do most ituexons terminate translation near their ends? It is known that premature termination codons (PTCs) upstream of the distal third of penultimate exons trigger transcript degradation while more 3′ PTCs fail to signal transcript targeting (13). Could this be a mechanism to prevent *bona fide* termination codons from being recognized as 'premature'?

A total of 410 5′UTRs and 432 3′UTRs (these numbers are different from those in exon size distributions because no definite boundary at the CDS end of the exon is required) were extracted in full length from 5utexons and 3tuexons, respectively. Their distributions and the quantile statistics are also plotted in Figure 3. In the human *GLA* gene for α-D-galactosidase A (X14448), the minimum 3′UTR (–2 nt in our definition) results from the fact that the CDS (including the stop codon) ends at 11 268 and the poly(A) site is at 11 266. UTR size distributions are useful, for example, when analyzing human expressed sequence tags (ESTs) and cDNA clones.

Mono- and dinucleotide compositions. Because compositional measures depend on the G+C content of the isochore (14) in which a gene is residing, for convenience, we compiled the statistics separately for data from low GC (<0.5) and high GC (≥0.5) genomic loci. The average G+C content of our data set is 0.53, but the average G+C content of all genomic loci from which our data set was extracted is only 0.49. The fundamental mono- and dinucleotide compositions in Table 1 are for the following groups of sequences: upstream flanks (216 nt), upstream UTRs, upstream introns (54 nt), CDS in each frame, downstream introns (54 nt), downstream UTRs and downstream flanks (216 nt). The values are represented as the percentage difference relative to the average of the total data set.

In contrast to the averages, some of the salient features are given below. At low GC loci, the average G+C contents of the various genomic regions have the following order: *uutr*>*uflk*> average *cds*>*din*>*dflk*>*dutr*>*uin* (see Table 1 for the notations). A codon has a consensus of RWY (in IUPAC ambiguity codes). At high GC loci, the average G+C contents of intron elements are boosted in such a way that the new order becomes:

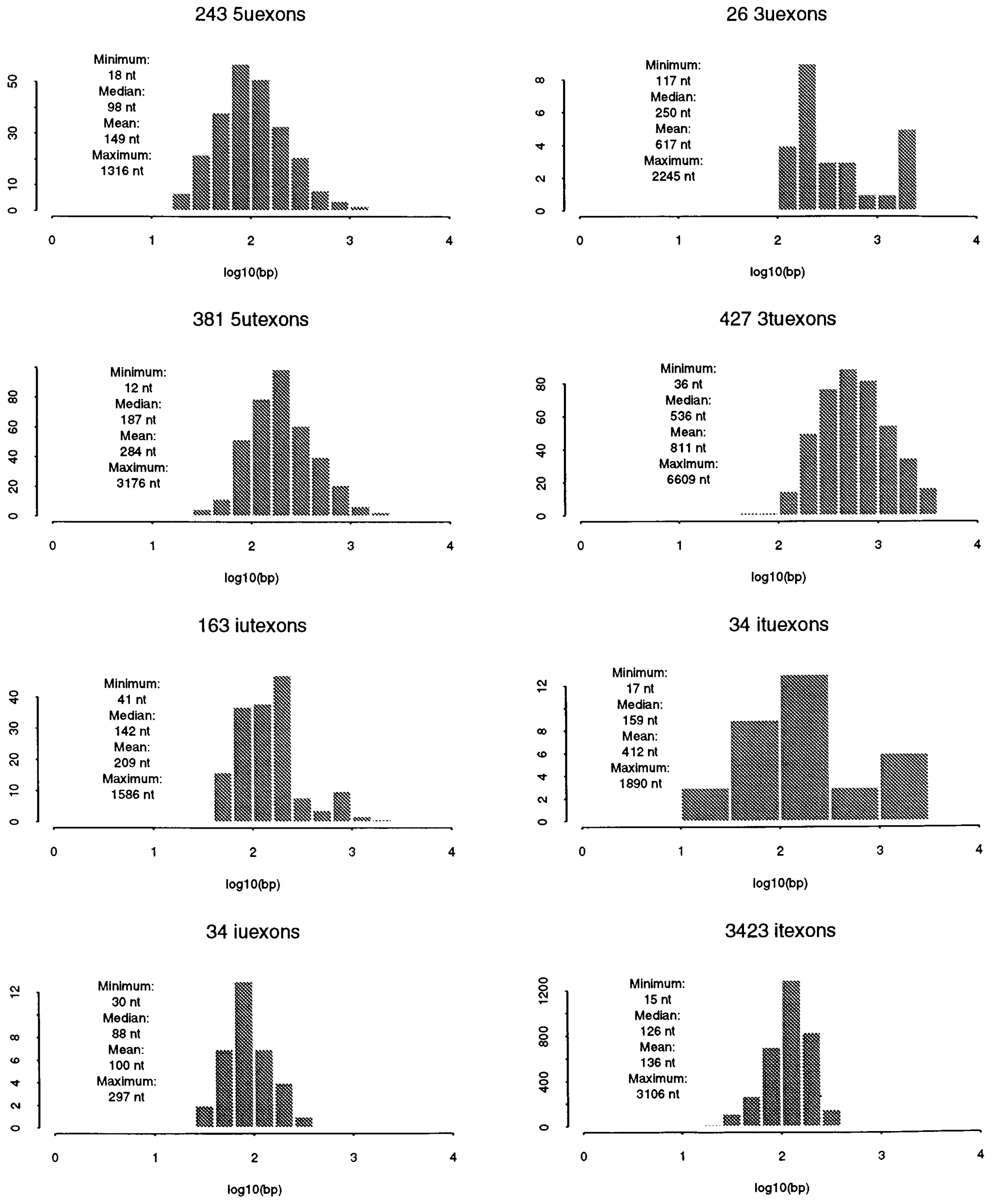

Figure 2. Exon size distributions for the first eight categories (4731 exons from GenBank: 87.0). The quantile statistics were calculated in log10 (bp) scale and converted back into units of nucleotides for ease of comparison.

uutr>*uflk*>*din*> average *cds*>*uin*>*dflk*>*dutr* (notice the dramatic increase of G in *din* and of C in *uin* both at the expense of T); a codon has a different consensus of SWS. At the dinucleotide level, the lack of self-complementary pairs is quite obvious. At low GC loci, they are CG and GC (CG is especially scarce because of the methylation decay); at high GC loci, they are TA and AT. However, they are relatively enriched at the 5′ (for CG and GC) and the 3′ end (for TA and AT) of a gene where they may play some role in the control signals of transcription (e.g. the CpG islands are often found near the 5′ end of a housekeeping gene, and the AATAAA motifs are often found near the polyadenylation site at the 3′ end of a gene. In *uin*s, the richness of dipyrimidines is clearly caused by the poly(Y) splicing signal near the 3′ splice site (3′ss), and the C content is controlled by the

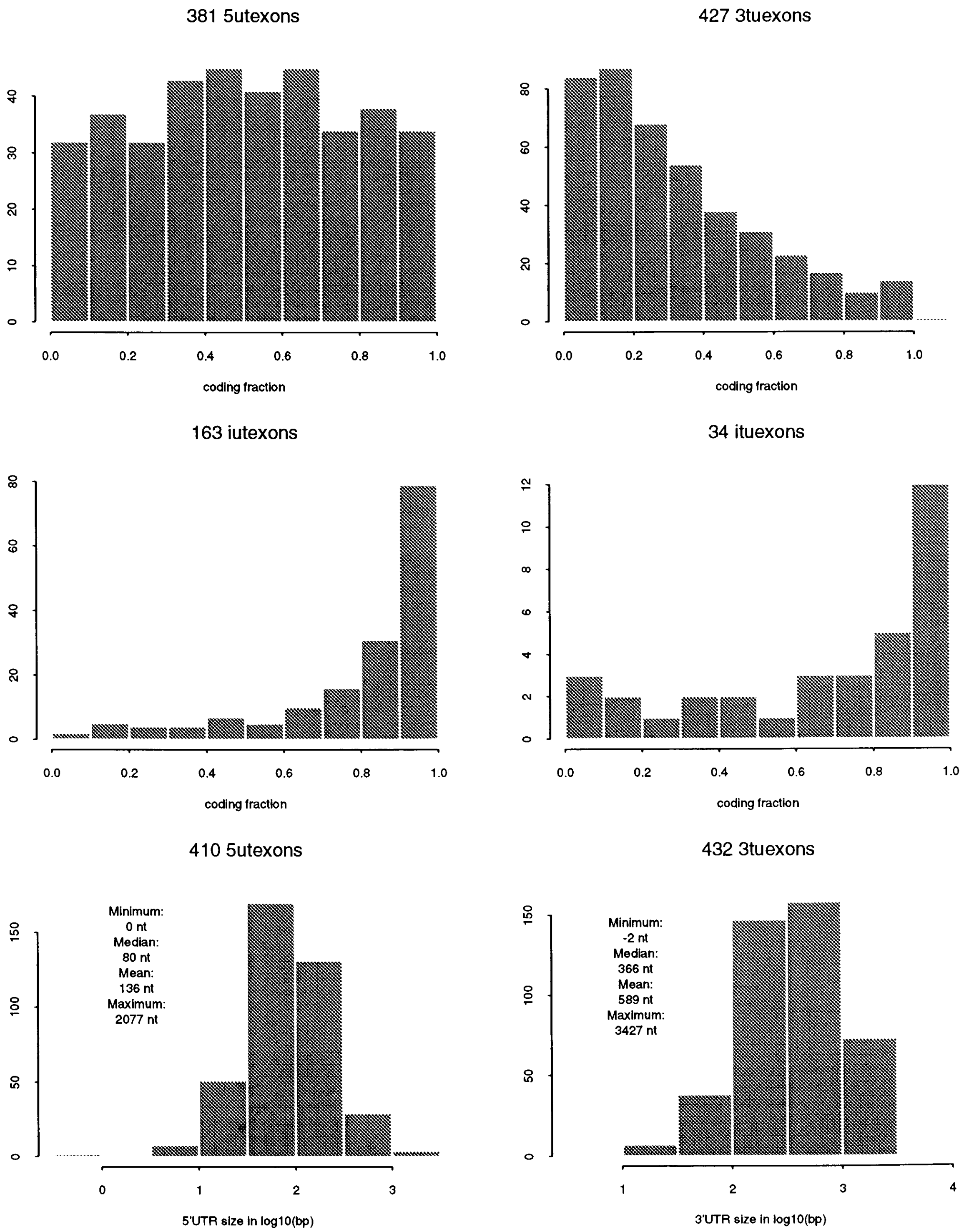

Figure 3. Coding fraction (top two panels) and UTR distributions (bottom panel). The coding fraction is defined by the ratio of the size of the coding portion over the total size of an exon. The quantile statistics for the UTRs were calculated in log10 (bp) scale and converted back into units of nucleotide.

G+C level of the genomic loci. Also, the poorness of AG in *uin* may be caused by avoiding the confusion of the true 3′ss. Unexpectedly, the richness of GG in *din* (which is related to the G-string excess, see later discussions) is only associated with high GC loci. Because of the coding constraints, the dinucleotide bias in CDS is strongly correlated with the codon bias below (see also 15).

Codon usage. The trinucleotide statistics are presented as in the human codon usage table (Table 2). The most abundant codons

Table 1. Mono- and dinucleotide compostions for low/high G+C loci expressed as percentage difference from the average (of the entire category 1–8 data set)

Nucleotide Compositions (N=1.7Mb)
(relative to the total average, in percent)

$G+C<0.5$	uflk	uutr	uin	cds1	cds2	cds3	din	dutr	dflk	ave
a	-1.7	-3.2	-5.1	1.2	5.1	-4.4	0.5	1.8	0.0	27.8
c	2.6	5.5	-0.5	-0.3	0.9	4.4	-2.9	-2.4	-2.4	21.4
g	3.2	3.5	-6.6	9.5	-3.0	1.9	0.0	-2.8	-1.3	21.8
t	-4.1	-5.8	12.1	-10.3	-3.0	-1.9	2.4	3.4	3.7	29.0
aa	-0.4	-1.8	-2.0	1.9	-0.2	-2.4	0.3	1.3	0.0	9.0
ac	-0.3	0.2	-0.6	0.6	1.9	-0.9	-0.7	-0.2	-0.3	5.0
ag	0.5	1.7	-3.4	-1.6	2.4	2.2	0.5	-0.7	-0.5	6.9
at	-1.5	-3.0	0.9	0.2	1.0	-3.4	0.4	1.3	0.8	7.0
ca	-0.1	0.5	-1.9	-0.1	-0.2	3.3	-1.2	-0.3	-0.3	6.9
cc	1.6	2.1	-0.2	-0.2	1.2	1.0	-1.1	-1.0	-1.2	5.7
cg	1.4	2.1	-0.9	0.8	-0.2	1.0	-0.4	-0.8	-0.8	1.6
ct	-0.3	0.8	2.4	-0.7	0.1	-0.6	-0.2	-0.3	-0.1	7.3
ga	-0.2	0.1	-2.7	5.4	-1.8	0.8	-0.4	-0.9	-0.9	6.4
gc	1.4	2.7	-1.6	1.6	0.7	0.1	-0.7	-0.9	-0.9	4.7
gg	2.3	1.3	-2.6	1.7	-1.3	2.2	0.3	-1.1	-0.1	5.4
gt	-0.2	-0.7	0.3	0.9	-0.6	-1.6	0.9	0.2	0.7	5.2
ta	-0.7	-2.0	1.1	-1.9	-2.0	-0.6	0.9	1.8	1.4	5.4
tc	-0.3	0.6	2.6	-1.1	0.5	-0.6	-0.2	-0.4	-0.1	6.2
tg	-1.0	-1.4	-0.1	-3.9	0.9	4.4	-0.4	-0.1	0.1	7.8
tt	-2.2	-3.0	8.5	-3.4	-2.5	-4.9	1.9	2.0	2.2	9.7
#	46kb	29kb	77kb	98kb	98kb	98kb	70kb	193kb	30kb	740kb

$G+C\geq 0.5$	uflk	uutr	uin	cds1	cds2	cds3	din	dutr	dflk	ave
a	-1.5	-2.4	-4.9	3.6	9.0	-8.2	-1.3	2.1	1.4	19.7
c	1.1	3.5	5.1	-2.9	-6.2	8.1	-2.8	-2.1	-5.3	30.5
g	2.9	2.0	-6.6	4.7	-8.2	5.6	6.0	-4.0	-0.4	29.2
t	-2.5	-3.2	6.3	-5.3	5.4	-5.4	-2.0	3.9	4.3	20.6
aa	0.0	-0.7	-1.7	3.6	-0.8	-2.3	-0.8	1.4	1.0	4.1
ac	-0.9	-0.5	0.0	0.3	4.0	-2.4	-1.0	-0.1	-0.6	5.0
ag	0.0	0.6	-2.8	-2.9	5.4	-1.5	1.4	-0.3	0.3	7.3
at	-0.5	-1.5	-0.3	2.6	0.3	-2.2	-0.8	1.0	0.6	3.3
ca	-0.8	-0.8	-1.0	-0.2	-2.7	5.2	-0.8	0.2	-0.2	7.0
cc	0.9	1.5	3.9	-4.0	0.3	0.7	-0.2	-0.5	-3.0	11.0
cg	2.3	3.3	-2.0	0.3	-0.7	2.4	-1.1	-2.1	-2.3	4.2
ct	-1.3	-0.5	4.2	0.9	-3.1	0.2	-0.6	0.3	0.1	8.3
ga	-0.6	-0.4	-2.2	5.0	-3.6	1.2	0.4	-0.6	-0.1	6.3
gc	1.4	2.5	-1.6	-0.5	0.6	1.5	-0.1	-1.9	-2.0	8.4
gg	2.5	0.4	-2.6	-1.5	-3.6	3.3	5.1	-1.6	0.5	9.8
gt	-0.4	-0.6	-0.1	1.7	-1.6	-0.7	0.7	0.3	1.3	4.6
ta	0.2	-0.6	-0.2	0.7	-1.0	-0.6	-0.4	1.3	0.7	2.1
tc	-0.5	0.1	3.8	-2.1	2.9	-2.9	-1.0	0.2	0.1	6.3
tg	-1.8	-2.1	0.3	-4.0	4.6	0.8	0.3	0.1	1.2	7.8
tt	-0.4	-0.5	2.3	0.1	-1.1	-2.7	-0.9	2.3	2.3	4.4
#	86kb	75kb	105kb	137kb	137kb	137kb	103kb	151kb	30kb	961kb

uflk/*dflk*, up-/downstream flanking region (216 nt); *uutr*/*dutr*, up-/downstream untranslated region; *uin*/*din*, 5′/3′ splice site intron region (54 nt); *cds*1–3: coding sequence in frame 1–3.

(CUG, GAR and AAG) are related to dominant dinucleotides in the first coding frame (GA, AA for genes at low GC loci, and CU, GA for genes at high GC loci). The least abundant codons (NUA and NCG) are related to the rare dinucleotide at the second position (CG for genes at low GC loci and UA for genes at high GC loci). There are other examples, such as the GCC codon for alanine and the CAG codon for glutamine, that have strong isochore bias. The most abundant amino acids (leucine and serine) are the ones that have most codons and have no rare dinucleotide at the first position. Stop codon UAA is particularly avoided for genes at high GC loci, although all three stop codons are used equally for genes at low GC loci.

Hexamer (6-tuple) statistics. Hexamer frequency has been used widely as a major discriminant factor in exon/intron identification (10,16). We have calculated hexamer frequencies f_{exon} (from all the CDSs in all frames) and f_{intron} (from the introns of 43 complete sequenced genes). The ratio f_{exon}/f_{intron} for human is less discriminating than in the fission yeast (11) because the human splice sites are more degenerate (see the splicing signals later). In Figure 4, the dot charts of some extreme ranking frequency differences are depicted separately for low and high G+C loci. Most of the intron characteristics can be explained by the run of Ts or As. However, at high GC loci, the presence of CA repeats and G-strings is apparent, especially the complementary pair GGGAGG/CCTCCC, which occurs much more often in introns. In the coding regions, the information is dominated by the hexamers consisting of frequent codons (especially in tandem repeat) that are also highly biased by the G+C content. Both codon usage and hexamer statistics are very useful, for example, when choosing appropriate restriction enzymes or designing various probes.

Table 2. Human codon usage table for low/high G+C loci (in percent)

Phe(F)	UUU	2.2/1.0	Ser(S)	UCU	1.8/0.9	Tyr(Y)	UAU	1.8/0.7	Cys(C)	UGU	1.3/0.7
	UUC	1.8/2.5		UCC	1.7/2.0		UAC	1.5/2.0		UGC	1.2/1.6
Leu(L)	UUA	0.8/0.2		UCA	1.4/0.7	Ter(*)	UAA	0.1/0.0	Ter(*)	UGA	0.1/0.1
	UUG	1.4/0.8		UCG	0.3/0.6	Ter(*)	UAG	0.1/0.1	Trp(W)	UGG	1.3/1.3
Leu(L)	CUU	1.4/0.8	Pro(P)	CCU	1.8/1.7	His(H)	CAU	1.2/0.6	Arg(R)	CGU	0.5/0.5
	CUC	1.5/2.4		CCC	1.4/2.8		CAC	1.1/1.6		CGC	0.6/1.8
	CUA	0.7/0.4		CCA	1.9/1.4	Gln(Q)	CAA	1.6/0.7		CGA	0.6/0.6
	CUG	2.8/5.6		CCG	0.3/1.0		CAG	2.8/4.0		CGG	0.6/1.6
Ile(I)	AUU	2.1/0.9	Thr(T)	ACU	1.6/0.9	Asn(N)	AAU	2.3/0.9	Ser(S)	AGU	1.3/0.7
	AUC	2.0/2.6		ACC	1.8/2.5		AAC	2.0/2.2		AGC	1.6/2.1
	AUA	0.9/0.2		ACA	1.7/1.0	Lys(K)	AAA	3.2/1.1	Arg(R)	AGA	1.4/0.5
Met(M)	AUG	2.2/2.2		ACG	0.4/0.9		AAG	3.4/3.5		AGG	1.0/1.0
Val(V)	GUU	1.5/0.5	Ala(A)	GCU	2.2/1.7	Asp(D)	GAU	2.7/1.4	Gly(G)	GGU	1.4/1.1
	GUC	1.4/1.8		GCC	2.0/3.9		GAC	2.2/3.2		GGC	1.9/3.5
	GUA	0.9/0.3		GCA	1.7/1.2	Glu(E)	GAA	3.9/1.5		GGA	2.5/1.5
	GUG	2.3/3.7		GCG	0.4/1.1		GAG	3.1/5.2		GGG	1.3/2.2

Total number of codons: 98 392/136 953 from low/high G+C sequences.

Signal profiles

All the signal profiles reported below are essential in delineating human gene organization. They also provide information about factor-binding energies around recognition sites (17).

Promoter signals. Human promoter sequences are difficult to identify and are poorly annotated in the public databases. We used a refining procedure to obtain the various promoter signal profiles (Table 3), where the vertebrate scores of Bucher (4) were used as the initial approximate matrices (their positional coordinates were also adopted). The windows and the number of sequences used in the search are shown, together with the mean distance of the signals from the transcriptional start (CAAT and GC boxes were searched on both strands). This simple refining procedure was able to produce reasonable promoter signal profiles consistent with the well-known consensus (4). Of course, genes that have no TATA or CAAT box would have contributed 'noise' to the profile frequencies (such noise could be reduced by imposing a minimum score requirement, as was done for the branch site profile below). As mentioned in (4), we also found a 1 nt shift in the human Cap-site annotations. Measuring from the end of a box, the TATA box is found ~25 nt and the CAAT box ~100 nt (with larger deviations) upstream from the transcriptional start site.

Pre-mRNA 3′ end processing signals. The AATAAA box and the poly(A) site profiles are shown in Table 4. The first AATAAA box profile was obtained by aligning all of the annotated poly(A) signals. The second was obtained from a subset of the signals that occur within a 50 nt window upstream of the poly(A) site. The second profile allowed us to estimate the mean distance (16 nt) between the T in the box to the poly(A) site. The first poly(A) site profile was also obtained simply from aligning all of the annotated poly(A) sites. Due to the uncertainty in identifying a precise poly(A) site, we do not see the CA consensus from this profile. In an attempt to correct some possible errors, we re-aligned the sequences if there was a CA within a ±2 nt distance from the annotated poly(A) site. We believe this poly(A) site profile (shown at the bottom of the table) may be closer to the truth.

Translational signals. Both the translational start and stop profiles were obtained by aligning the corresponding sequences according to the annotations (to avoid possible errors, only the consensus sequences starting with ATG or ending with one of the three stop codons were compiled; Table 5). The human translational start profile is consistent with the general consensus for all vertebrates: GCCGCCRCCATGG (18). From 531 sequences, we also found that start codons occurred as the first, second, third or fourth ATG in the open reading frame (ORF) 474, 51, five and one times, respectively [which would be in favor of the 'first ATG rule' and is very similar to our previous finding for the fission yeast (11)]. The translational stop profile shows the ratio among different stop codons as TAA:TAG:TGA ~1:1:2 (obtained from TAA+TAG ~TGA and TAA+TGA ~3 TAG according to the matrix at positions –2 and –1), which may also be seen from Table 2.

Splicing signals. The 5′ and 3′ splice site profiles were obtained by aligning annotated sequences obeying the GT–AG rule (Table 6). We also did separate profiles for the splice sites adjacent to UTR and those adjacent to CDS; we did not find any substantial differences (data not shown). From the general mononucleotide compositional analysis above, it was shown that the compositional property in the flanking intron regions depends strongly on the G+C content. One could get more insight by comparing splicing signal profiles for low and high G+C content; Table 6 is such a comparison calculated from itexons. Although the human splice site consensus more or less agrees with the general consensus for most vertebrates, AG|GTRAGT for the 5′ss and

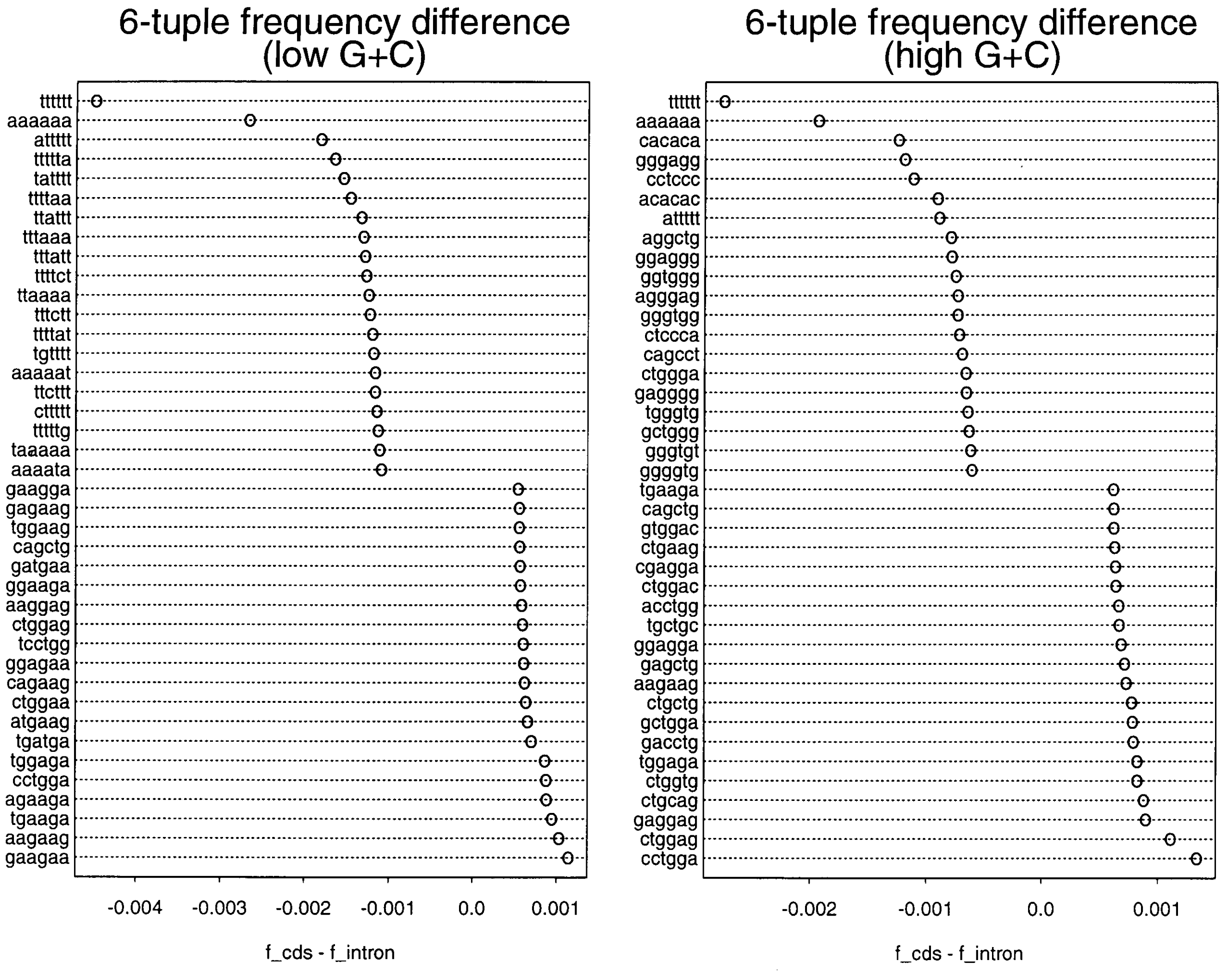

Figure 4. Major differences in 6-tuple frequency between coding sequences (in all frames) f_{cds} and introns f_{intron}. Only the top and the bottom 20 6-tuples are displayed.

(Y)nNCAG|G for the 3′ss, the G+C content greatly affects the nature of purine or pyrimidine constituents. At low G+C loci, the 5′ss consensus may be better described as AG|GTAAGT and the 3′ss consensus as (Y)nNYAG|G.

The branch site profile was again obtained by the refining method above, where the vertebrate scores (19) were used as the initial approximate matrix. We took a window (of size 41 nt) 10 nt upstream of the 3′ss end where most reported branch points were found (20). To reduce the noise created by genes that have their branch point located outside of the window, we imposed a minimum score of 3, corresponding to the 1st quantile of the maximum score distribution (the absolute conservation of A at the branch point was obtained automatically as a consequence). The average distance of the branch point from the 3′ss end was found to be 26 nt. Again the profile is biased by the G+C content. In contrast to the vertebrate consensus CTRAY, YTVAY for the low G+C content and CTSAY for the high G+C content may be more specific to the human sequence.

In addition to these conventional measures, we also examined many other statistics that may play a role in mRNA splicing. Some of these are plotted in Figure 5a–f (all the maximum values were limited by the searching window size used).

We plotted the distribution of the distance from the branch point to the 3′ss end (Fig. 5a); its quantile statistics are: Min. = 10, 1st Qu. = 20, Median = 25, Mean = 26, 3rd Qu. = 32 and Max. = 46. Most branch points are located 15–30 bp upstream of the 3′ss end. Branch points outside this region are suboptimal; this happens in many alternatively spliced introns (21).

We plotted the distribution of the distance from the 5′ss end to the closest (with respect to the CDS) downstream in-frame stop codon (Fig. 5b); its quantile statistics are: Min. = –2, 1st Qu. = 1, Median = 11, Mean = 16, 3rd Qu. = 27 and Max. = 51 (–2 occurs when the 5′ss looks like TA|GT where TAG is the first stop 'masked' by the 5′ ss boundary). The expected number of the first downstream stop codons found by chance in random sequences at each position is indicated by a dotted line. This is three times the expected number of the first in-frame downstream stop codons found by chance. The striking peak at 1 gives rise to the 5′ss consensus G|GTRAG, and there is a second peak at position 5 which may also correspond to a functional role in splicing. The fact that most significant stops are very close to the 5′ss supports the idea that the 5′ss may be there originally to mask the downstream nonsense codons (22) and to allow longer (and hence more complex) gene products to be coded. Many non-functional

Table 3. Promoter signal profiles (in percent)

Cap-site Profile: d=0.9(5), N=715, window=(-13,10)

	-2	-1	0	1	2	3	4	5
A	11	1	74	6	24	15	15	15
C	25	96	1	23	32	35	33	36
G	33	2	3	54	3	26	29	26
T	30	1	22	16	40	24	23	23
		C	A	G				

TATA-box Profile: d=27(8), N=738, window=(-50,-1)

	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
A	12	6	69	2	70	57	69	42	24	13	20	18	19	18	15
C	39	22	4	15	6	6	9	5	18	35	36	31	32	30	30
G	35	15	7	11	12	5	16	26	46	41	34	34	31	35	36
T	14	57	20	72	12	32	7	27	12	12	11	17	17	18	19
		T	A	T	A										

CAAT-box Profile: d=116(large), N=661, window=(-216,-30)

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4
A	24	16	20	50	23	3	2	91	48	15	15	50
C	41	33	33	4	11	92	95	4	12	3	46	11
G	12	26	20	41	53	4	2	2	27	25	35	35
T	23	24	27	5	12	1	1	3	13	57	4	4
						C	C	A	A	T		

GC(Sp1)-box: d=anywhere, N=693, window=(-164,1)

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	34	31	18	9	1	4	24	18	3	30	15	2	10	16
C	22	14	8	0	1	4	52	2	5	2	7	53	38	13
G	20	43	47	34	96	90	1	74	77	57	67	21	17	41
T	24	12	26	57	2	2	23	5	15	10	12	23	35	30
					G	G	C	G	G	G				

All these matrices were obtained by using the corresponding Bucher's matrices as the initial start in the refining procedure (see Database and Methods).

excess stops would have long ago been washed out by mutations in evolution.

We plotted the distance from the closest upstream AG to the 3′ss AG (Fig. 5c); its quantile statistics are: Min. = 2, 1st Qu. = 22, Median = 29, Mean = 30, 3nd Qu. = 37 and Max. = 52. We can see a sharp drop-off at short distances (<18 nt). Compared with Figure 5a, it is clear that the first AG downstream of the branch site is most likely to be used as the 3′ss. This rule works much better for the fission yeast (11).

To study more flanking intron features quantitatively, we looked at the Y-string in polypyrimidine [poly(Y)] tracts of the 3′ intron region and the possible G-string excess in the 5′ intron region.

The polypyrimidine tract is known to play an important role in human pre-mRNA splicing (23). A Y-string is a tandem stretch of pyrimidines. We extracted the maximum Y-string (closest to the 3′ss end) from the 54 nt upstream flanking intron region of 4417 exons and plotted the size and distance distributions in Figure 5d and e. The quantile statistics for the sizes are: Min. 2, 1st Qu. = 6, Median = 8, Mean = 9, 3rd Qu. = 11, Max. = 32; and for the distances are: Min. = 2, 1st Qu. = 4, Median = 9, Mean = 14, 3rd Qu. = 21, Max. = 50. Compared with the same statistics from random sequences having the same nucleotide composition (also shown in the figure), we see that the maximum Y-strings have a much longer length and occur very close to the 3′ss end in real introns (a *t*-test, df = 7899, showed the true mean 9.19 is outside of the 95% confidence interval of the random sample mean 6.62 ± 2.45). [A non parametric test (Wilcoxon rank sum) test (24) was also performed that definitely ruled out the null hypothesis that the medians of the two sample distributions are

Table 4. Pre-mRNA terminational processing signal profiles (in percent)

AATAAA-box Profile: N=540, all the annotated

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
A	32	32	28	30	33	34	96	87	6	91	97	96	37	30	26	22	22	25
C	19	19	21	20	24	31	2	1	0	1	1	1	17	19	22	22	21	21
G	18	17	17	18	16	13	1	2	1	0	1	1	24	20	19	14	16	15
T	31	32	34	32	27	23	1	10	92	7	2	1	22	32	33	41	41	39
							A	A	T	A	A	A						

AATAAA-box Profile: d=15(large), N=392, annotated in (-50,-1)

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
A	32	32	27	29	33	35	96	87	6	92	98	97	36	30	23	22	20	26
C	18	20	21	19	24	30	2	1	0	0	0	1	18	19	20	22	21	21
G	18	16	16	20	16	12	1	2	1	1	0	1	24	20	21	15	16	15
T	32	32	36	32	28	23	1	10	93	7	2	1	22	31	35	41	44	38
							A	A	T	A	A	A						

Poly(A)-site Profile: N=340

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
A	23	29	35	41	36	31	3	27	36	50	33	26	26	19	19	21	13	15
C	20	19	18	19	21	19	21	20	26	15	19	25	24	20	22	18	21	19
G	17	15	14	15	14	14	19	20	16	17	19	16	18	25	26	24	28	26
T	40	37	33	26	29	36	26	33	22	18	30	33	32	36	34	37	38	41
									a	a	a							

Poly(A)-site Profile: N=340, realign if a "CA" is within ±2 nt

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
A	25	29	34	38	36	32	36	29	22	64	29	24	28	21	17	17	16	16
C	21	18	18	19	20	20	15	16	46	8	15	26	22	21	23	17	19	20
G	15	18	14	15	15	13	20	21	13	13	24	16	19	24	24	28	26	25
T	39	35	33	28	29	35	29	34	19	15	33	34	32	33	36	39	38	39
									C	A	t							

The second and fourth matrices were obtained from the first and the third, respectively, by the refining procedures as described in the text.

Table 5. Translational signal profiles (in percent)

Start-site Profile: N=446, all the annotated

	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3
A	21	15	21	21	18	24	57	31	15	100	0	0	26
C	28	38	38	24	35	49	5	43	53	0	0	0	20
G	31	27	22	42	27	22	36	17	27	0	0	100	41
T	19	20	19	13	20	5	2	10	4	0	100	0	10
	g	c	c	g	c	C	A	c	C	A	T	G	

Stop-site Profile: N=641, all the annotated

	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9
A	26	37	15	0	51	76	28	27	21	16	23	19	21	25	20
C	29	22	40	0	0	0	19	24	27	32	30	30	29	27	31
G	26	15	29	0	49	24	38	32	28	27	23	25	26	24	26
T	20	26	16	100	0	0	14	17	23	23	23	26	22	22	22
		a	c	T	R	A	g	g							

Only sequences with the annotated start site ATG and the annotated stop site TAA, TAG or TGA were counted.

Table 6. Splicing signal profiles (in percent)

5'ss-site Profile: N=1394/1880

	-3	-2	-1	0	1	2	3	4	5
A	38/32	62/56	12/8	0	0	71/38	73/70	11/5	21/13
C	31/38	10/15	4/4	0	0	2/4	6/9	6/5	10/21
G	18/19	12/15	77/80	100	0	24/56	8/14	75/86	14/25
T	13/11	16/14	7/8	0	100	3/2	13/7	8/4	55/41
	n	A	G	G	T	R	A	G	t

3'ss-site Profile: N=1404/1891

	-15	-14	-13	-12	-11	-10	-9	-8	-7
A	15/10	14/ 8	13/ 7	11/ 8	10/ 6	10/ 6	11/ 4	12/ 8	13/ 8
C	24/41	21/42	20/41	22/40	21/38	22/43	25/42	28/46	28/49
G	10/15	12/14	10/14	9/13	10/13	9/12	10/13	10/14	8/10
T	51/34	53/36	57/38	58/39	59/43	59/39	54/41	50/32	51/33
	y	y	y	y	y	y	y	y	y

	-6	-5	-4	-3	-2	-1	0	1
A	11/ 7	10/ 6	26/19	7/ 2	100	0	26/21	24/19
C	25/54	22/45	25/38	55/82	0	0	11/13	15/21
G	5/ 8	5/ 8	15/26	1/ 0	0	100	50/58	20/29
T	59/31	63/41	33/17	37/16	0	0	13/ 8	41/31
	y	y	n	Y	A	G	g	t

Branch-site Profile: N=858/1128, d_{A-G}=26(10), window=(-50,-10)

	-5	-4	-3	-2	-1	0	1
A	25/15	25/15	0	0	39/16	100	18/7
C	19/36	22/38	60/88	2/5	24/31	0	33/56
G	15/24	17/22	0	0	32/51	0	3/12
T	41/25	36/25	40/12	98/95	5/2	0	46/25
	n	n	Y	T	V	A	y

For 5′ss and 3′ss matrices, only sequences that obey the standard GT–AG rule were counted. The banch site matrix was obtained via a refining procedure (see Database and Methods) using the Harris–Senapathy branch site matrix for vertebrates as the initial start.

the same which had a *P*-value <10^{-10}.] We expect the size effect would be larger if approximate Y-strings were used (allowing one mismatch, for example).

The G-string excess was hinted at in the above discussion of mono- and dinucleotide distributions. It was first reported by Solovyev *et al.* (25). We measured this feature in a more quantitative way as follows. Using itexons, for each minimum G-string size *i* (*i* = 1, 2, 3, 4, 5), we counted the number of tandem G runs of size *i* and larger in a 54 nt window on each side of the 5′ss boundary. Let the difference of the number on the intron side to the exon side be *D* and *d* = 1, 0 or –1 depending on whether *D* is >, = or < 0. The average *D* (unshaded) and *d* (shaded) are plotted in Figure 5f. One can see both measures peaked at 3, indicating that G triplets are the most over-represented G-strings on the intron side on average. Recent experiments indicated the G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection (26). It is possible that such G triplets may be related to hnRNP sites [such as hnRNP A1 sites which have a consensus of UAGGGU (27)].

CORRELATIONS AND CONSTRAINTS AMONG DIFFERENT SEQUENCE FEATURES

To explore possible relationships among different features across an exon, we used itexons that have complete flanking intron information. (For convenience of statistical tests, all variables in Figure 6a–d were normalized—subtracted the mean and divided by the standard deviation.)

Strong compensatory constraint between the two splice sites across a short exon. In Figure 6a, we have plotted the upstream 3′ss score against the downstream 5′ss score for each exon (we chose 0 as the minimum score cut-off, because the few splice sites with negative scores were mostly annotation errors). There appears to be no correlation for the whole data set at first sight. However, if we highlighted the exons that have a short size (< 61 nt) and have a relatively long upstream intron (>200 nt), there appeared to be a constraint (represented by the dotted line) that restricted these exons into the upper-right corner. To test the significance of the constraint, we found that the *P*-value (see Database and Methods) was 0.029 (*N* = 77, the constraint is $y = –2.8x – 3.5$). Therefore, this constraint is statistically significant at the 95% confidence level. Such a constraint is consistent with the Exon Definition model: both splice sites have to be consensual for a short exon not to be skipped and, when the 3′ss is weak in a long intron, a stronger downstream 5′ss is needed to compensate for exon definition. Experimentally, it was observed that mutation of a 5′ss depressed the removal of the upstream intron 20-fold (28). Strengthening a naturally weak 5′ss of an internal exon by making it a better fit to the consensus increased *in vitro* splicing of the upstream intron (29,30). *In vivo*, mutant 5′ splice sites usually were suppressed by second mutations that improved the 3′ss across the exon (31,32).

*Correlation between the splice sites and exon size.*To see how the minimum quality of splice sites depends on exon size, we made a box plot of the total splice site score (i.e. the sum of the 5′ss score and 3′ss score across an exon) as a function of the internal exon size (Fig. 6b, in log10). We observed that when the exon size is near the mean or larger, there appears to be no restriction on the splice site scores (except the absolute minimum); however, when the exon size decreases, the splice site scores tend to go up systematically. This is also consistent with Exon Definition in the sense that the interacting splicing factors across an exon may require an optimal interaction range. Experimentally, in addition to exon skipping, the other major phenotype resulting from mutation of a splice site in a human gene is activation of a cryptic site with the right polarity (33). When a constitutively recognized internal exon was internally deleted below 50 nt, it was skipped by the *in vivo* splicing

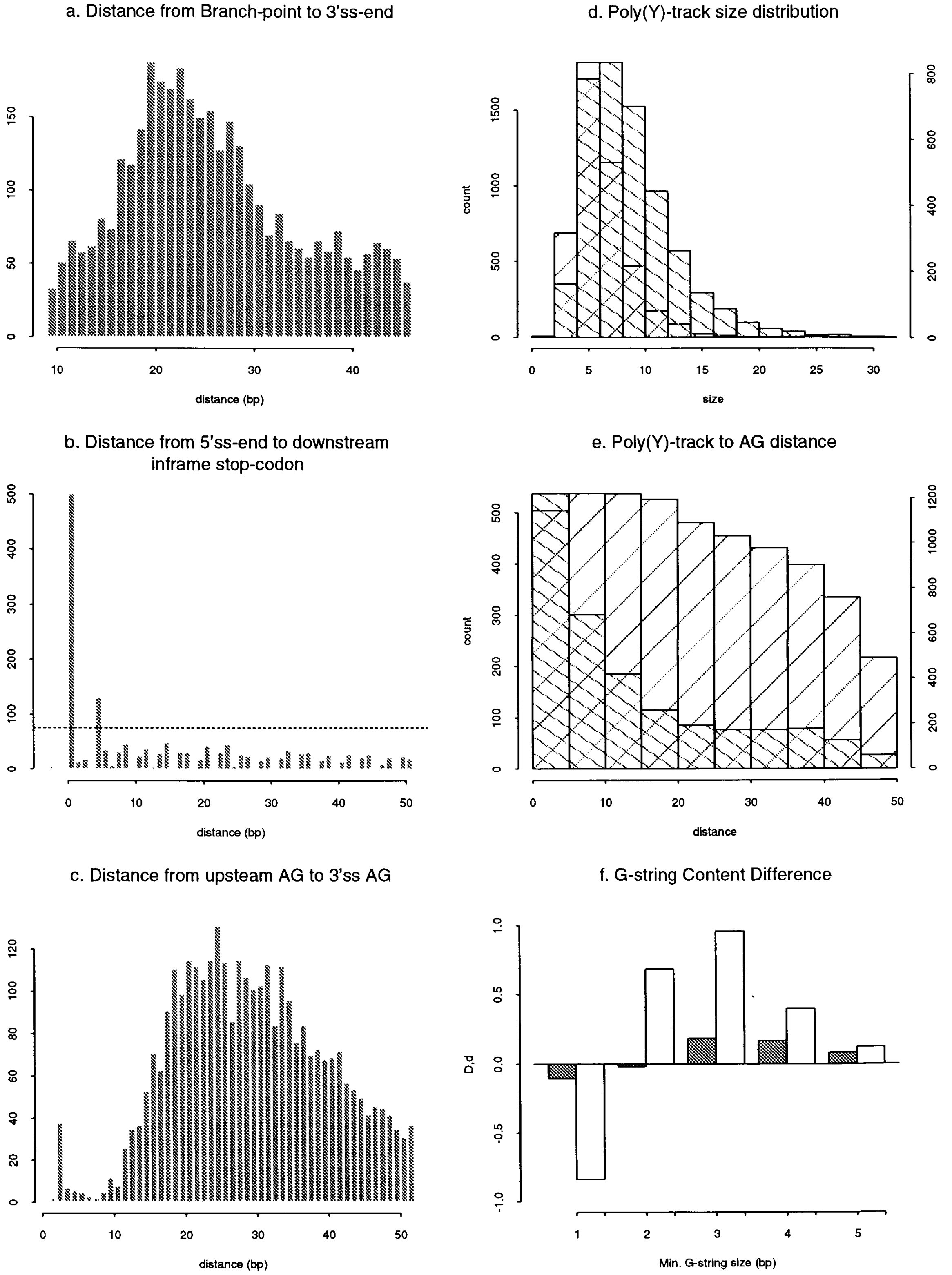

Figure 5. Novel splice site features. (**a**) The distance from the branch point to the 3′ss is mostly between 15 and 30 bp. (**b**) The distance from the 5′ss to the first downstream in-frame nonsense codon is mostly 1 bp. The dotted line indicates the expected number of the first downstream nonsense codon which is three times the expected number of the first in-frame downstream nonsense codon. (**c**) The distance from the 3′ss to the first upstream AG is mostly between 18 and 35 bp. (**d**) The poly(Y) tract in the 3′ sequence of real introns is larger than would be found by random chance. (The histogram for the random model with the same nucleotide composition is cross-hatched.) (**e**) The distance from the poly(Y) tract to the 3′ss is much shorter than would be found by chance. (The histogram for the random model is cross-hatched.) (**f**) G-string excess in the 5′ss intron region is mostly G-triplets. *D* and *d* are two different statistical measures (see text). The peak at 3 bp indicates that G-triplets are more enriched in the 5′ss intron region relative to the adjacent upstream exon region.

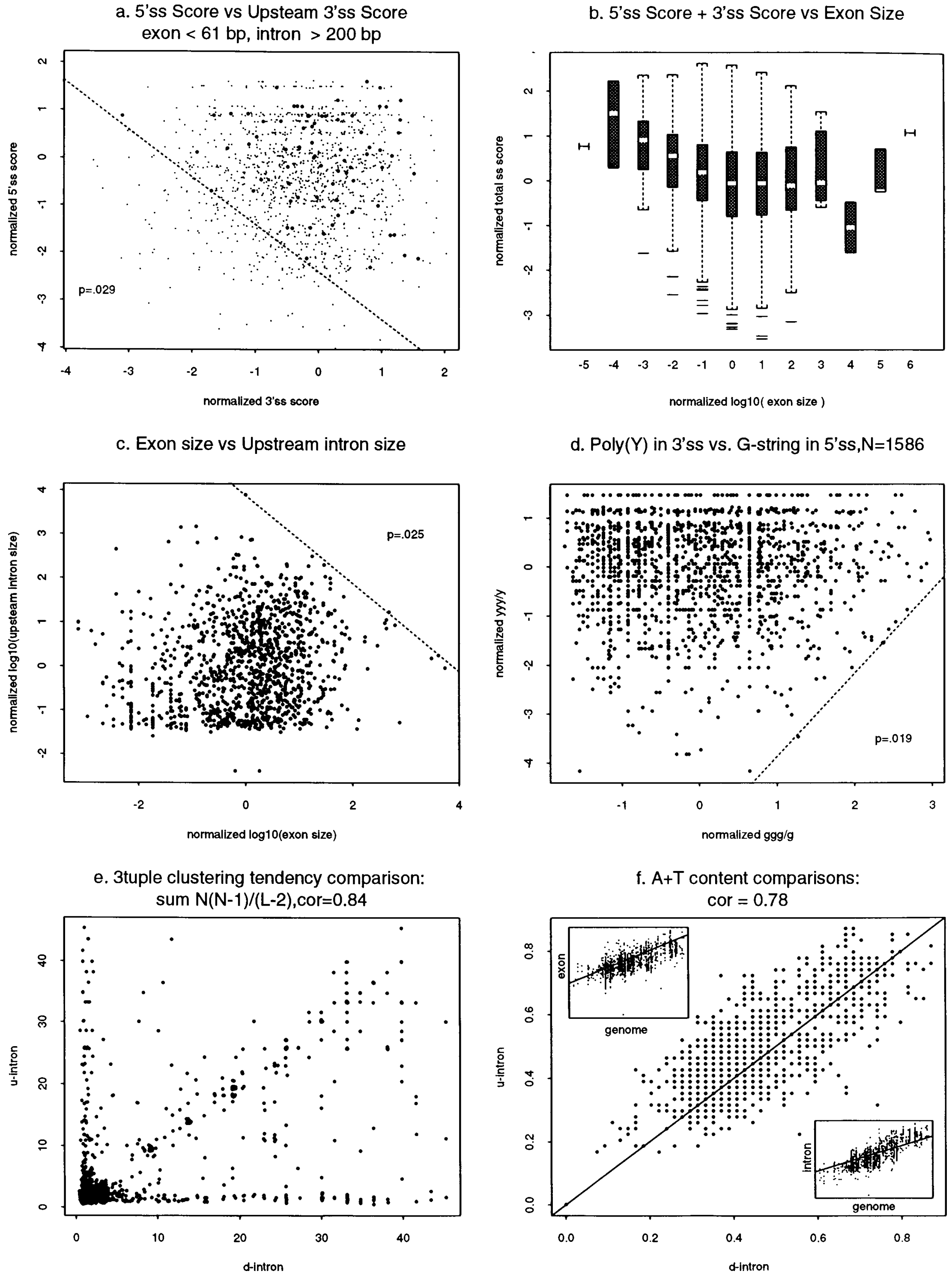

Figure 6. Correlations and constraints among different sequence features. (**a**) Scatter plot of 5′ss scores versus 3′ss scores (both normalized, see text). The highlighted dots represent the short exons (<61 bp and flanked by an upstream intron >200 bp). That fact that these highlighted dots are bounded by the dotted line indicates the compensatory constraint between minimum qualities of upstream 3′ss and downstream 5′ss across these short exons. (**b**) Correlation between the splice site quality (measured by the sum of the flanking splice site scores) and the exon size (normalized in log10 scale) is apparent for short exons. (**c**) The constraint on minimum upstream intron size for large exons is indicated by the dotted line. (**d**) The constraint between the upstream poly(Y) (measured by the normalized ratio of YYY frequency to Y frequency) and downstream G-strings (measured by the normalized ratio of GGG frequency to G frequency) is indicated by the dotted line. (**e**) A 3-tuple clustering tendency correlation is indicated by the diagonal stripe in the scatter-plot of the 3-tuple clustering tendency (which measures how likely it is that one would find mononucleotide runs) in the upstream flanking intron region versus the downstream flanking intron region. (**f**) Scatter-plots of A+T contents for the upstream flanking intron region versus the downstream flanking intron region, for the exon region versus the whole genomic region and for the total flanking intron region versus the whole genomic region are displayed. The A+T content correlations are indicated by the diagonal stripes.

machinery (34). Increasing the strength of the splice sites could revert the mutant (35).

Constraint between the upstream intron size and exon size. To prevent steric hindrance between splicing factors, a minimum intron size is generally expected. This is the case, as may be seen in Figure 6c where the upstream intron size is plotted against the exon size (both in log10). With our data set, this minimum was 24 nt (for the two introns in the human parvalbumin gene); the others were 60 nt or larger. Unexpectedly, our data also show a restriction at the upper-right corner. Namely, a long exon is often accompanied by a short upstream intron. A significance test arrived at $P = 0.025$ ($N = 1228$, the constraint is $y = -x + 3.88$), implying that the constraint is significant at the 95% level. According to the Exon Definition, when the exon size becomes very large, the upstream intron may only be recognized through the Intron Definition mode (33), which would require a shorter intron size (see also ref. 36).

Constraint between the poly(Y) tract and the downstream G-strings. To demonstrate that the G-string feature mentioned above is also correlated with the poly(Y) tract, we designed a simple measure as follows: for a G-string measure, we counted the total size of G-strings (G runs of 3 nt or more) in a 54 nt window downstream of a 5′ss and normalized it by dividing the total number of G residues; we constructed a similar measure for the poly(Y) tract by counting the total size of Y-strings (pyrimidine runs of 3 nt or more) in a 30 nt window upstream of the 3′ss and by dividing the total number of pyrimidine residues. As shown by the dotted line in Figure 6d, the relationship manifests itself again as a constraint: when the G-string content downstream of an itexon is too high (>1 on the normalized scale), the poly(Y) tract upstream has to be of better quality (more Y-triplets clustering). Again the features on the opposite sides of an itexon appear to be 'talking' to each other. The P-value for the constraint is 0.019 ($N = 1586$ and the line is $y = 1.7x - 5.58$), again significant at the 95% level. It is possible that the G-strings bind the Y-strings across an itexon transiently to help initial exon recognition during RNA splicing.

Correlation of 3-tuple clusterings in the flanking introns. Due to runs of As or Ts (as shown, for example, in Fig. 4), most introns have less complexity than exons (37). We wanted to see if there are any correlations between clusterings in different regions. We use $x = P_k N_k(N_k - 1)/L - 2$ as our 3-tuple clustering measure (this measure has been used by Roman Tatusov in his low complexity filtering software—dust) where N_k is the number of a 3-tuple k, L is the window length and the sum is over all possible (64) 3-tuples. The larger this number is, the stronger is the 3-tuple clustering. We had calculated x for each side of the 3′ or 5′ splice site with a window of length 54 nt; we observed that most of the introns have lower clustering tendency than the exons, and there were no correlations between different exon regions or between an exon region and an intron region (data not shown). However, there was a strong correlation between the upstream intron region and the downstream intron region for a subset of itexon data (as indicated by the diagonal stripe in Fig. 6e). The correlation coefficient was *cor* = 0.84 with a 90% confidence interval of (0.825,0.846) (defined in Database and Methods).

Correlation between flanking A+T contents across an itexon. Finally, we show a correlation between mononucleotide compositions in the two flanking intron regions (54 nt each) across an itexon. We have calculated A+T content in each of the following regions: upstream flanking intron, itexon and downstream flanking intron, and compared them with the A+T content of their genomic locus. We found that the correlation between flanking introns across an itexon was the strongest, with *cor* = 0.78 and a 90% confidence interval of (0.758,0.802) (as seen in Fig. 6f). Presumably this correlation is caused mainly by the isochore effect (14), as may be seen from the subplots in Figure 5f, which shows that genomic A+T content is clustered into islands and the flanking intron A+T content correlates with the genomic more positively than the exon A+T content (the straight lines are the equal A+T content lines). This correlation is remarkably strong because only very short (54 nt) flanking intron sequences were used and the poly(Y) or the G-string in these regions would have to adjust its composition to accommodate the correlation. In fact, all the signal profiles are GC content-dependent as shown earlier. That is why exon identification in a low GC locus is very difficult as the exon GC content is constrained to be almost as low as introns and intergenic regions.

CONCLUSIONS

In summary, we have analyzed statistical characteristics of many sequence features in human exons and their flanking regions, including some very subtle and complex ones. We have begun to reveal several novel correlations and constraints among different features. Some of these are in accord with recent experimental observations; others are still a mystery awaiting functional interpretation. We should emphasize that most correlations are between extreme values (mathematically, the relationships can only be expressed as inequality constraints). For features that are close to their consensus, they are quite free to vary and the exon will still be defined (recognized). This type of degeneracy is quite typical for biological systems (e.g. non-lethal mutations of protein sequences are often tolerated because the resulting change of a structure is not critical to the function). However, under special conditions where the general constraints are violated, the exons have to be defined by a delicate balance of multiple features and/or a requirement for additional new features [such as the secondary structures (38), purine-rich enhancers (39), etc.]. These have been observed often in alternatively spliced genes (40). All correlations related to gene structures are likely to be complex, because they resulted from dynamic interactions of many macromolecules during evolution. However, these interdependencies among different features are just as important as each individual feature. This is analogous to the fact that protein function cannot be worked out by structure characterization alone without considering interactions between subcomponents or with substrates. More rigorously characterizing existing features, further discovering new features, and quantitatively exploring novel feature relationships will be the keys for understanding gene structures and for improving exon discrimination methods. Many statistical findings reported here have already been utilized in a new gene-finding method (41) which has substantially improved the accuracy in human exon prediction. Better understanding of the architecture of genes will become the prerequisite for innovative experimental designs in functional studies.

DATABASE AND METHODS

Human exons (in nuclear protein-coding genes) were extracted from GenBank release 87.0, and the data set was processed to remove redundant copies and checked for data integrity. Starting with 31 202 human sequences extracted out of gbpri.seq, we filtered out viruses, mitochondria, RNAs, pseudogenes, Ig genes, MHC genes, redundant large family genes and identical copies. Only 6152 sequences remained. Of these, 780 contained complete CDSs which belong to the last four categories (see Fig. 1). We sorted the rest in ascending order according to their size, and extracted all the exons that had no >90% maximum similarity to other larger exons. Many errors were removed or corrected during the entire analysis by comparison against the original publications. The final data set consisted of 5061 exons (representing 2705 intron-containing genes) belonging to the first eight categories (see Fig. 1). These exons and their flanking regions (54 nt into introns or 210 nt into intergenic regions) have been deposited in FASTA format at the anonymous ftp site phage.cshl.org in the directory pub/science/human_exons. The accession number and the coding frame information are also retained in each exon record.

All profiles were defined by positional dependent frequency matrices (17)

$$f_{\alpha x} = n_{\alpha x}/N, \text{ with } N = \sum_{\alpha} n_{\alpha x},$$

where $n_{\alpha x}$ is the counts of a k-tuple α at position x. The scores were defined by

$$s_{\alpha x} = \log_2 (P_{\alpha x}/P_0),$$

where $P_0 = (1/4)^k$ (if α is a k-tuple) and $P_{\alpha x}$ isthe Bayesian posterior probability (42) given by

$$P_{\alpha x} = (n_{\alpha x} + a_{\alpha})/(N + A), \ with\ A = \sum_{\alpha} a_{\alpha}$$

We chose the pseudocount a_α to be $\sqrt{N}$ multiplied by the background frequency of α. The counts were obtained either from the known alignments or by the following refining procedure. Starting with an approximate matrix (we took the corresponding vertebrate matrix; one may also take a matrix resulting from a known alignment of subset data), align the signals (within a specified window) that have the maximum score and calculate the new matrix. One then iterates this procedure until it converges.

To assess a straight line constraint in a plan, we used the normal approximation (with a mean of zero and standard deviation of unity) by normalizing (subtracting the mean and dividing by the standard deviation) the two sample variables x_i and y_i. (The variables may need to be transformed, such as by taking the logarithm of exon sizes, so that their distribution can be approximated by the normal distribution.) Assuming x, y are independent, the null hypothesis is that the observed N points are not restricted by a straight line (generalization to an arbitrary curve is straight forward) $y = ax + b$. The probability f of finding one point in the restricted region is given by $F(d)$, where F is the cumulative distribution function and $|b|/\sqrt{1 + a^2}$ is the distance of the constraint line from the origin. The P-value for finding all N points on one side of the constraint is f^N.

The definition of the correlation coefficient *cor* of two vectors of sample data X,Y, is the standard:

$$cor = Ave. \frac{(X - \mu_1)(Y - \mu_2)}{\sigma_1 \sigma_2},$$

where μ_1, μ_2 and σ_1, σ_2 are the means and standard deviations, respectively, of X and Y. The confidence interval for *cor* is obtained by the procedure described in (43).

Most of the analysis and graphics were done with SPLUS (44).

ACKNOWLEDGEMENTS

This work was supported by NIH grant HG00010. The author is grateful for referees' detailed suggestions on how to improve the manuscript.

REFERENCES

1. Breathnach, R. and Chambon, P. (1981) Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.*, **50**, 349–383.
2. Mount, S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.*, **10**, 459–472.
3. Senapathy, P., Shapiro, M. and Harris, N.L. (1990) *Methods Enzymol.*, **183**, 252–278.
4. Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
5. Hawkins, J.D. (1988) A survey on intron and exon lengths. *Nucleic Acids Res.*, **16**, 9893–908.
6. Smith, M.W. (1988) Structure of vertebrate genes: a statistical analysis implicating selection. *J. Mol. Evol.*, **27**, 45–55.
7. Stormo, G.D. (1988) Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Chem.*, **17**, 241–263.
8. Staden, R. (1990) Finding protein coding regions in genomic sequences. *Methods Enzymol.*, **183**, 163–180.
9. Gelfand, M.S. (1990) Global methods for the computer prediction of protein-coding regions in nucleotide sequences. *Biotechnol. Software*, **7**, 3–11.
10. Fickett, J.W. and Tung, C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
11. Zhang, M.Q. and Marr, T.G. (1994) Fission yeast gene structure and recognition. *Nucleic Acids Res.*, **22**, 1750–1759.
12. Robberson, B.L., Cote, G.J. and Berget, S.M. (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.*, **10**, 84–94.
13. Belgrader, P. and Maquat, L.E. (1994) Nonsense but not missense mutations can decrease the abundance of nuclear mRNA for the mouse major urinary protein, while both types of mutations can facilitate exon skipping. *Mol. Cell Biol.*, **14**, 6326–6336.
14. Bernardi, G., Mouchiroud, D. and Gautier C. (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.*, **28**, 7–18.
15. Karlin, S. and Mrazek, J. (1996) What drives codon choices in human genes? *J. Mol. Biol.*, **262**, 459–472.
16. Claverie, J.-M., Sauvaget, I. and Bougueleret, L. (1990) K-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods Enzymol.*, **183**, 237–252.
17. Stormo, G. D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–221.
18. Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.
19. Harris, N.L. and Senapathy, P. (1990) Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucleic Acids Res.*, **18**, 3015–3019.
20. Green, M.R. (1991) Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu. Rev. Cell Biol.*, **7**, 559–599.
21. Helfman, D.M. and Ricci, W.M. (1989) Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Res.*, **17**, 5633–5650.
22. Senapathy, P. (1988) Possible evolution of splice-junction signals in eukaryotic genes from stop codons. *Proc. Natl Acad. Sci. USA*, **85**, 1129–1133.
23. Coolidge, C.J., Seely, R.J. and Patton, J.G. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.*, **25**, 888–896.

24. Lehmann, E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks*. Holden and Day, San Francisco, CA.
25. Solovyev, V.V., Salamov, A.A. and Lawrence, C. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res*., **22**, 5156–5163.
26. Carlo, T., Sterner, D.A. and Berget, S.M. (1996) An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *RNA*, **2**, 342–353.
27. Chabot, B., Blanchette, M., Lapierre, I. and La Branche, H. (1997) An intron element modulating 5′ splice site selection in the hnRNP A1 pre-mRNA interacts with hnRNP A1. *Mol. Cell. Biol*., **17**, 1776–1786.
28. Talerico, M. and Berget, S.M. (1990) Effect of 5′ splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol*., **10**, 6299–6305.
29. Kuo, H.-C., Nasim, F.H. and Grabowski, P.J. (1991) Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. *Science*, **251**, 1045–1050.
30. Grabowski, P.J., Nasim, F.H., Kuo, H.-C. and Burch, R. (1991) Combinatorial splicing of exon pairs by two-site binding of U1 small nuclear ribonucleoprotein particle. *Mol. Cell. Biol*., **11**, 5919–5928.
31. Carothers, A.M., Urlaub, G., Grunberger, D. and Chasin, L. (1993) Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. *Mol. Cell. Biol*., **13**, 5085–5098.
32. Tsukahara, T., Casciato, C. and Helfman, D.M. (1994) Alternative splicing of beta-tropomyosin pre-mRNA: multiple *cis*-elements can contribute to the use of the 5′- and 3′-splice sites of the nonmuscle/smooth muscle exon 6. *Nucleic Acids Res*., **22**, 2318–2325.
33. Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem*., **270**, 2411–2414.
34. Dominski, Z. and Kole, R. (1991) Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol*., **11**, 6075–6083.
35. Dominski, Z. and Kole, R. (1992) Cooperation of pre-mRNA sequence elements in splice site selection. *Mol. Cell. Biol*., **12**, 2108–2114.
36. Sterner, D.A., Carlo, T. and Berget, S.M. (1996) Architectural limits on split genes. *Proc. Natl Acad. Sci. USA*, **93**, 15081–15085.
37. Konopka, A.K. and Owens, J. (1990) Complexity charts can be used to map functional domains in DNA. *Gene Anal. Technol. Appl*., **7**, 35–38.
38. Clouet d'Orval, B., d'Aubenton Carafa, Y., Sirand-Pugnet, P., Gallego, M., Brody, E. and Marie, J. (1991) RNA secondary structure repression of a muscle-specific exon in HeLa cell nuclear extracts. *Science*, **252**, 1823–1828.
39. Tanaka, K., Watakabe, A. and Shimura, Y. (1994) Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol*., **14**, 1347–1354.
40. Stamm, S., Zhang, M.Q., Mar, T.G. and Helfman, D.M. (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res*., **22**, 1515–1526.
41. Zhang, M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.
42. Tanner M. and Wong W.H. (1987) The calculation of posterior distribution by data augmentation. *J. Am. Stat. Assoc*., **82**, 528.
43. Snedecor, G.W. and Cochran, W.G. (1980) *Statistical Methods.* 7th edn, Iowa State University Press, Ames, IA.
44. S-PLUS User's Manual (1991) Vol. 1, Chap. 6, Sept. Statistical Sciences, Inc., Seattle, WA.