1750–1759 Nucleic Acids Research, 1994, Vol. 22, No. 9

Fission yeast gene structure and recognition

M.Q.Zhang* and T.G.Marr
Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, USA

Received August 16, 1993; Revised and Accepted March 9, 1994

ABSTRACT

A database of 210 *Schizosaccharomyces pombe* DNA sequences (524,794 bp) was extracted from GenBank (release number 81.0) and examined by a number of methods in order to characterize statistical features of these sequences that might serve as signals or constraints for messenger RNA splicing. The statistical information compiled includes splicing signal (donor, acceptor and branch site) profiles, translational initiation start profile, exon/intron length distributions, ORF distribution, CDS size distribution, codon usage table, and 6-tuple distribution. The information content of the various signals are also presented. A rule-based interactive computer program for finding introns called INTRON.PLOT has been developed and was used to successfully analyze 7 newly sequenced genes.

INTRODUCTION

More than 15 years have passed since the discovery of the 'amazing' [1] process of precursor messenger RNA (pre-mRNA) splicing, in which introns are removed and exons are joined to form mature mRNA (see [2] for recent reviews). This process is carried out in a spliceosome, composed of U1, U2, U4–U6, and U5 small nuclear ribonucleoproteins (snRNP's) and numerous non-snRNP protein factors. Indeed, the spliceosome cycle [2] starts with formation of the Commitment Complex in which the binding of U1 to the donor site is essential; with the help of U1 snRNP, U2 snRNP joints to form the stable Complex A by binding to the branch site; spliceosome assembly is completed when the tri-snRNP U4/U6.U5 enters to form Complex B1 in which U5 is base paired with the acceptor site; freed from an inhibitory base-pairing with U4, the same region of U6 base pairs with U2 to induce an important conformational change from Complex B1 into B2; the first transesterification step — cleavage at the 5′ splice site and formation of the lariat intermediate mark the formation of Complex C1; the second transesterification step — cleavage at the 3′ splice site and ligation of the exons result in Complex C2 (the two reactions have been recently shown to occur at two distinct catalytic sites (3)); finally the spliceosome is disassembled in order to allow snRNPs recycling, after releasing the spliced exons for transporting out of nucleus and excised intron lariat for degradation. It is evident that assembly of the trans-acting factors into a spliceosome is directed in part by highly conserved pre-mRNA sequences at the donor, acceptor and branch sites [4]. In addition to these short splicing signals, exon/intron sizes and compositions also play important roles [5].

Recent understanding of the splicing process has come from powerful genetic studies of the budding yeast (*Saccharomyces cerevisiae*) and of the fission yeast (*Schizosaccharomyces pombe*) [6]. Although the two yeasts are equally amenable to genetic and biochemical analysis, their introns are distinctly different in characteristics [9]. For example, budding yeast introns are much larger (like mammalian introns); but the splicing signals of the budding yeast are much more strongly conserved (unlike mammalian introns). About 10% of the genes in the budding yeast contain introns, most of which are found in ribosomal protein genes [10]. In the fission yeast, introns appear to be widespread. Roughly 44% of the genomic sequences surveyed contained introns, and the average number of introns per complex gene ('complex' genes contain introns, 'simple' genes do not) is about 2 (little more than half of the complex genes contain multiple introns ranging from 2 to 7). It has been shown that insertion of artificial introns into naturally intronless *S.pombe* gene leads to an efficiently spliced gene producing a functional product [7, 11], which suggests that the proper signals within an intron under normal circumstances are sufficient to initiate and complete splicing.

Having completed a 13 kb resolution physical map sequencing project [12], We began our gene recognition study in preparation for the Cold Spring Harbor fission yeast. We have completed an analysis of 163 *S.pombe* genomic sequences using a number of computational methods. Our study extends the previous survey [9] of 36 fission yeast genes and our methods are similar to those used in the studies of *Caenorhabditis elegans* and *Drosophila* [13]. Current gene-finding programs (Gm [14], GRAIL [15], GeneId [16], SORFIND [171, GeneParser [18], to name a few) can be divided in two classes: those that use a classical statistical approach and those that use a neural-net approach (or the combination of the two). The two classes of methods are equivalent in principle, since they both have to calculate signal profile matrices, various length constraints, ORFs and codon bias from the existing data set. This information is used to summarize the data into rules and criteria to make predictions regarding a gene structure. Due to the ambiguities in the gene start and end signals (promoters, enhancers and poly(A) signals) of a pre-mRNA transcript, none of the existing software can predict the first and last exon reliably from a large genomic DNA sequence.

* To whom correspondence should be addressed

As a first step towards an accurate and automatic gene finding program, we developed an interactive intron-finding program by assuming the approximate positions of the ends of a gene are known. Our program has been successfully used to help biologists at our laboratory to find the coding regions in newly sequenced genes. In these genes, the clone fragments were relatively short and it was known that each fragment contained a functional gene. The main feature that distinguishes our approach from all the others is the incorporation of two additional informations: the branch site profile and the minimum distance constraint (between a branchpoint and its downstream acceptor), and the ability to identify short introns (< 50 bp).

DATABASE AND METHODS

Data sets

Database entries containing *S.pombe* DNA sequences were taken from GenBank release 81.0. The data set was manually edited to remove redundant copies, tRNA, rRNA and snRNA genes, mitochondrial DNA sequences, partial genes, histones, centromere, telomere and other repetitive or questionable sequences. Data were also verified by checking the original publication. As shown in Table 1, there are 210 genes (180 are genomic and 30 are of cDNAs) used in the statistical analysis. Among the genomic sequences, 101 contain simple genes and 79 contain complex genes. The complex genes have a total of 174 introns. There are an additional 4 complex genes which were collected privately and used in intron recognition tests. Some of the statistics were computed after we divided the 174 introns into two classes: short introns (less or equal to 200 nt) and long introns (the rest) [7,8,19]. We did this because it has been shown that the extension of short introns with random sequences over 200 nts leads to very poor splicing efficiency or the complete abolishment of splicing [7] (this cutoff length is twice as long as that used in some previous works [13,11]).

Methodology

All the profiles are presented by frequency matrices [20]. Information contents are defined and computed as in [23]. Size statistics are calculated and plotted by S-PLUS [21] functions. Since these signals are located at different boundaries of statistically-distinct regions, the background frequency of each base is set to .25 as was done in [11]. The intron finder — INTRON.PLOT is written in the S-PLUS script language, the exon/intron content discrimination is based on 6-tuple distributions [24]. Precise mapping of transcriptional initiation/termination sites has been extremely difficult, due to the degeneracies of promoter, enhancers and poly(A) signals. Our analysis of first/last exons have been restricted to first/last coding exons which are defined as those with translational initiation/termination sites. Identification of translational initiation/termination sites and coding exon/intron junctions is one of the main objectives of our gene recognition software.

RESULTS AND DISCUSSION

Information contents and profiles

The information content $I(x)$ of [23] reflects the extent of deviation from random base composition at position x. It is important to estimate how strong the translational start and splicing signals are by studying their information content.

As shown in Fig. 1, there is a peak in information content at each of the boundary signals, corresponding to the traditional consensus sequences. The width of each peak indicates the extent of each signal. The corresponding profiles for the following signals are also tabulated in the figure:

Translational initiation start. The fission yeast translational start signal is similar to the budding yeast (consensus 5′-(A/Y)A(A/U)AAUGU-3′) and the higher eukaryotes (consensus 5′-CACCAUGG-3′) both with regard to the conserved AUG codon and an 'A' nucleotide at the −3 position. This is the most conserved nucleotide among different eukaryotes, which is consistent with the finding [25] that this A nucleotide is more important to maintain than C nucleotides at the other positions to achieve efficient utilization of an AUG codon in higher eukaryotes. The upstream A-richness makes the fission yeast appear to more closely resemble to the budding yeast in this regard, but the slightly preferred G at +4 position is more typical of higher eukaryotes. This observation supports the notion that these two yeasts have diverged from each other just as much as each has diverged from human [26].

According to the scanning mechanism (see [27] for a recent review), the position of an AUG codon, relative to the 5′-end of the transcript is as important as its context. Without adequate information about leader sequences of the fission yeast, we have examined which AUG in the first translated ORF is actually used. There are only four (out of 210) pre-mRNAs where the first AUG is not used as the start (the second AUG is used instead). The four genes are: (a) *gsa1*, the first AUG site is AUUAUAUGG which is 216 nt upstream of the start site GCAAAAUGA; (b) *cdc8*, the first is UAUGCAUGU which is 63 nt upstream of the start GGAAAAUGG; (c) *rhp6*, the first is AGUGAAUGG which is 138 upstream of the start UAAAAAUGU; (d) *sod2*, the first is AAGGCAUGU which is 222 nt upstream of the start UACUAUGG. It is striking that none of these 4 first AUG sites has an A nucleotide at −3 position, while the second (the start) AUG sites do have (they apparently also have better overall consensus). This should be compared with the finding in the survey [19] that 95% of mRNAs follow first AUG rule both for the budding yeast and higher eukaryotes, and at the −3 position, the relative frequencies are 75% A; 13% G; 8% C; 3% U for the budding yeast and 79% A; 18% G; 3% Y for higher eukaryotes, while at +4 position, the relative frequencies are 39% U; 29% A; 24% G; 8% C for the budding yeast and 39% G; 29% A; 17% C; 16% U for higher eukaryotes.

Donor site. The donor site is the 5′ splice site located at the exon/intron boundary. We examined the 9-nucleotides in the consensus region [28]. It was known that all of the naturally occurring thalassemia mutations in human, β-globin gene occurred within this region [29] and a G to U mutation at the +1 position resulted in an invariant exon skipping in the human α-galactosidase A causing Fabry disease [30].

When the donor site profile was computed separately for short/long introns, the difference between donor site profiles in short and long introns was not much (data not shown), although the donor site profile in long introns does appear to be closer to the consensus AAG|GUA(A/U)GU. The sequence C-AG|GUAAGU has been thought of as the prototype donor sequence, because it is 100% complementary to a sequence at the 5′ end of the U1 snRNA in mammals [31]. It has been suggested that during splicing, the 5′ sequence of the U1 snRNA

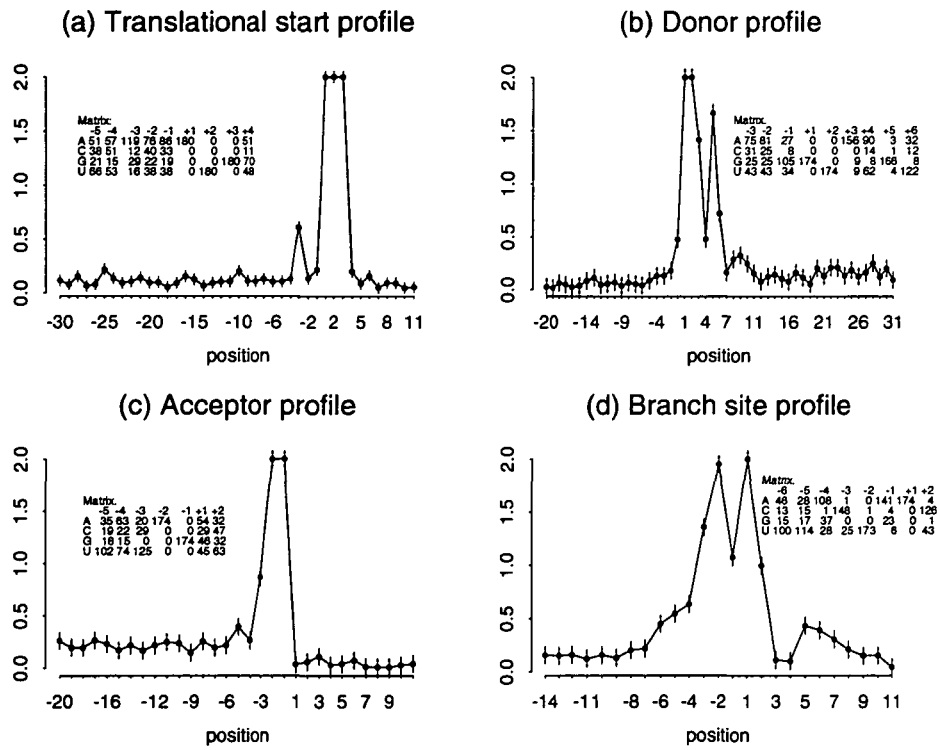

Figure 1. Information contents and profiles for boundary signals: (a) translational start site (the upstream leader sequences for *pab1*, *pim1* and *spi1* genes were not not available); (b) splice donor site; (c) splice acceptor site and (d) branch site.

pairs with the donor sequence [32]. Analysis showed that the U1 snRNA from *S.pombe* displays a sequence identical to the mammalian one at the 5′ end [33], but we only found one sequence (in intron I of *prh1*) with a perfect match:

5′- *C A G |G U A A G U-3′* *prh*1
3′- *G U C C A U U C A ppGm*3 *U*1.

Almost all the donor sequences have a minimum of 5 matches (out of the 9-nucleotides) to U1 as suggested in [9], except four (*ck1:I, chk1:V, cut7, nda3:I*) having 4 matches and one (*ypt1:I*) having only 3 matches. Some of these could be due to annotation errors. In the budding yeast introns, GUAUGU is highly prevalent [34], since together with GUAAGU they are found frequently in other lower eukaryotic introns and relatively frequently in plant, insect and vertebrate introns [35,31]. It is conjectured [9] that one of these may be the ancestral donor site. We want to point out that the absence of the C nucleotide at the position +3 may be important for splicing in the fission yeast.

Acceptor site. The acceptor site is the 3′ splice site located at the intron/exon boundary. The consensus (U/A)UAG can only be seen on the intron side. Again, the absence of G nucleotide at the position −3 implies all important constraint. There were no significant differences between acceptor sites in short and long introns (data not shown). Comparing with the consensus sequences of AYAG from the budding yeast [8], GYAG from plants [35] and NCAG from mammals [36], we see they share only the YAG as the common signal. It was proposed recently [4] that, through a Holliday structure, U1 snRNA may base-pair with the intron sides of both donor and acceptor sites and U5 snRNA basepairs with the exon sides of both. The conserved AG acceptor end is required for specific binding with U1 snRNA.

Branch site. In mammalian introns the sequence YNCURAC has been identified as the consensus branch sequence [37]. In the budding yeast, the sequence UACUAAC is highly conserved at the branch site [8]. Since most of the branch sites of *S.pombe* were not mapped, we chose to use motif identification software — RTIDE [22] to identify putative branch sites. We aligned all the introns by the acceptors and selected a 35 nt window immediately upstream of the last, 3 nucleotides at the 3′ ends of the introns. With the setting of word size 6, 2 mismatches and no insertions/deletions, RTIDE found ACUAAC as the consensus word which had the highest score of 106.83 and highlighted all the words of size 6 which had at most 2 mismatches (see Figure 2).

After studying some mapped branch sites, we used the following hierarchy of rules to pick the putative branch sites: (a) mapped branches; (b) higher scoring site; (c) closest to 3′ end if there is a tie. From these putative branch sites, we calculated the profile as shown in Figure 1c.

Comparing the less stringent consensus (U/A)AYURAY with the mammalian and the budding yeast consensus sequences, we see they share YURAY as the common branch motif. It has been proposed that a sequence in the U2 snRNA base pairs with the branch site sequence UACUAAC in the budding yeast, bulging out the last A nucleotide [8]. This argument has been extended to mammalian systems. The branch site binding sequence of U2 is absolutely conserved in both mammals and the fission yeast [38]. The basepair interaction between the consensus branch site and U2 snRNA is indicated below:

```
                    A
5'- U|A  A   Y   U   R       Y -3'      branch consensus
3'-  A   U   G   A   U   —   G UG-5'    U2.
```

Size distributions

Internal intron length properties. It is known that the 3′ splice region in fission introns is often very short and that it does not appear to have polyuridine tracks [9]. We have plotted the distribution of the distance between branch point A and the 3′ conserved dinucleotide AG in Fig. 3; these distances are ranged from 5 to 24 with 9 being the median. The main feature is that the majority of 3′ splice sites tends to be short but there exits a constraint on the minimum distance of 5 nt.

To examine possible correlation between intron length and the C+U content, we made scatterplots of C+U content vs. length separately for the region between the donor GU and the branch point A and the region between the branch point A and the acceptor AG (data not shown). It does not show any significant correlations, which suggests that a polyuridine track is not required in the splicing of *S.pombe* introns. Therefore, some of the splicing factors may not be necessary in this organism.

U run characteristics between long and short introns. There is no evidence of ploy(U) tracks in the 3′ region and the few long introns (19 out of 174 introns have size larger than 200 nt) are characterized by their long 5′ region. It was reported [9] that there exists poly-U islands in the region between the donor and branch sites of long *S.pombe* introns. These are U runs often interrupted by C or G, but very rarely by an A.

We first looked at the statistics of the longest U runs in this region. We made a scatterplot (data not shown) of the length of the maximum U run found in this 5′ region of each intron against the natural logarithm of the length of the region. If the U's were distributed randomly on larger scale, then, according to the Erdös and Rényi law [39], the expected maximum U run should be increasing linearly on that scale, when the length of the 5′ region increases (the proportional constant is $log(1/P(U))$, where $P(U)$ is the composition of U). The LOWESS smoothing (a robust, local smoothing algorithm, see [40]) showed an approximate agreement with the asymptotic theory, which indicated there are no significant long pure poly(U) tracks even in the long 5′ regions (data not shown). To investigate shorter and impure U runs, we made a 4-tuple frequency plot of short 5′ region against long 5′ region (data not shown). Indeed, the interrupted shorter U runs are characteristic of introns. 'UUUU' had the highest frequency, its single nucleotide substitutions by A or G are the second highest. Furthermore, the substitution by G is more likely in long 5′ regions and by A in short 5′ regions. Concatenations of these 4-tuples could explain, to some extent, the previous observations regarding poly-U islands. These signals are not powerful enough to discriminate long introns from short ones.

Intron size distributions. In Fig. 4a, we have plotted the histogram of all 174 introns in the 79 complex genes. The median size of all the fission yeast introns was about 63 bp (1.8 on a log10 scale) it should be compared with 79 bp of *Drosophila* [13] and 600 bp of Vertebrate and is much shorter in comparison with plants and mammals. Short introns are also characteristic in most fungi [41], except in the budding yeast where the majority of introns are between 250 and 550 bp [34]. Although most fission yeast

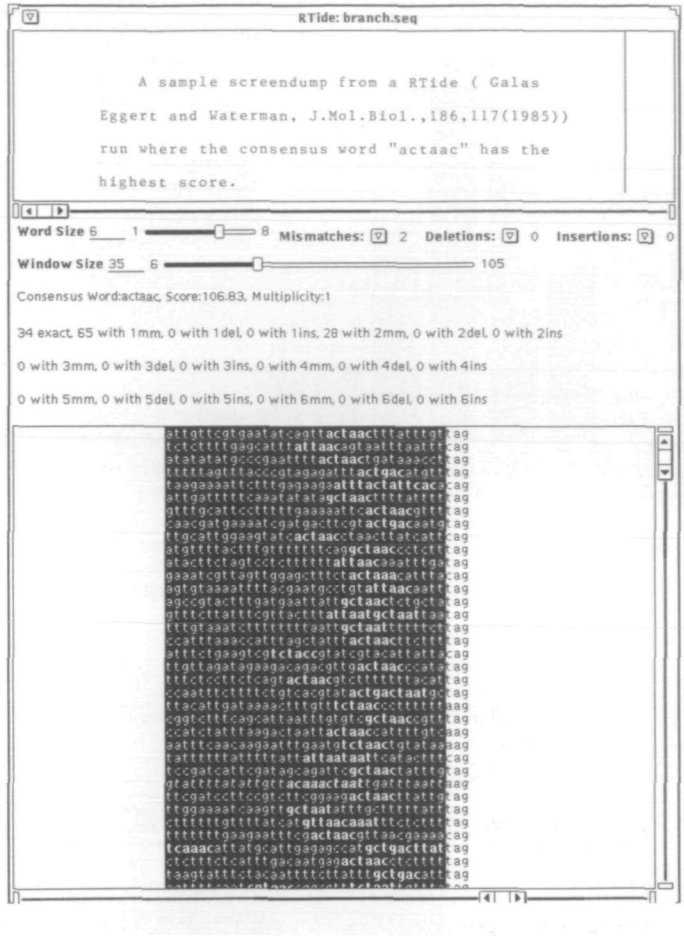

Figure 2. A sample screendump from a RTIDE run.

introns are short, there are also some long ones as shown in the tail of the distribution. The shortest (*cut7*:I or *rpb1*:I) is 35 bp and the longest (*dbp2*:II) is 699 bp. When the data are divided in three groups: (1) introns in single intron genes; (2) longest introns in multi-intron genes and (3) the rest; it shows that the size of the longest introns in the multi-intron genes peaks around 2.1 (126 bp), while the rest in the multi-intron genes are very short. The sizes of the introns in the single intron genes are more variable, even though the average length is about the same as that of the longest introns in the multi-intron genes (Fig. 4a).

We also grouped and analyzed the introns from the 30 multi-introns genes into (1) the first intron; (2) the intervening introns (3) the last intron (Fig. 4b). It can be seen that the size distribution of all the introns (Fig. 4a) is about the same as that of the introns in the multi-intron genes (Fig. 4b). But among the latter, the intervening introns tend to be short.

Coding exon size distributions. The coding exons in 180 genes were divided into three groups: (1) simple genes; (2) the longest in complex genes and (3) the other coding exons in complex genes (Fig. 4c). The total median length is about 336 bp (this figure is not readily comparable with the median length 133 bp of Vertebrate exons [5], because more than half of the fission yeast sequences contained no introns) The shortest of the coding exons in published sequences was 3 bp (*atp*a:I and *cam1*:I). We found a even shorter one in a new gene from our laboratory, see the

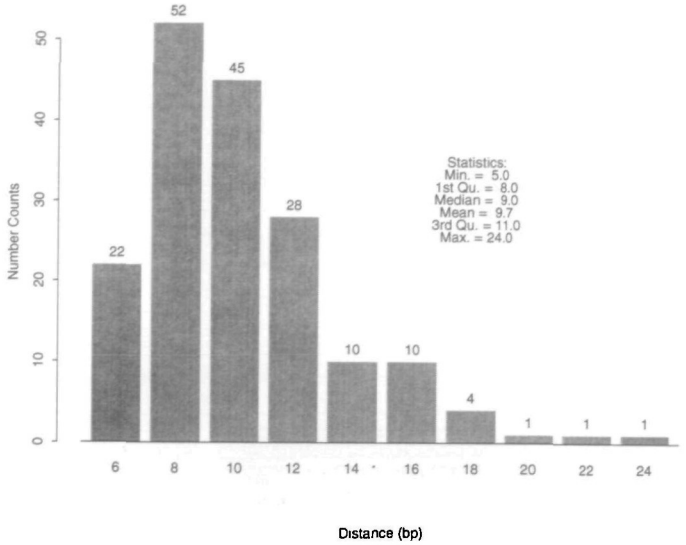

Figure 3. Distribution of distance between branch point A and acceptor AG in 174 *S.pombe* introns.

last section). the longest exon was 5079 bp (*cyr1*). Groups (1) and (2) have approximately the same size distribution, centered about 3.2 (1585 bp) and group (3) is centered about 2.1 (126 bp, which is comparable with the 133 bp size of Vertebrates). The log-normal shape of the exon size distributions is striking in contrast with the intron size distributions.

We grouped and analyzed the coding exons in 61 complex (intron containing) genes into: (1) first exons; (2) the intervening exons and (3) last exons (Fig. 4d). The 'first' and the 'intervening' exons had similar distribution centers around 2.2 (158 bp) and the 'last' had a center at 3.1 (1259 bp). In most complex genes, the shortest exon was the first and the longest one was the last.

ORF length distribution in the 180 genomic DNA sequences. In Figure 5, we plot the distribution of the ORF (Open Reading Frame) sizes in both strands of the 180 genomic sequences. The majority of them are very short with a median of 29 bp (1.48 on a log10 scale). The size distributions of the short ORFs are the same in both strands. The longer ORFs are biased towards the sense strands because most of these comprise the coding regions. All the longest ORFs in sense strands are coding and,

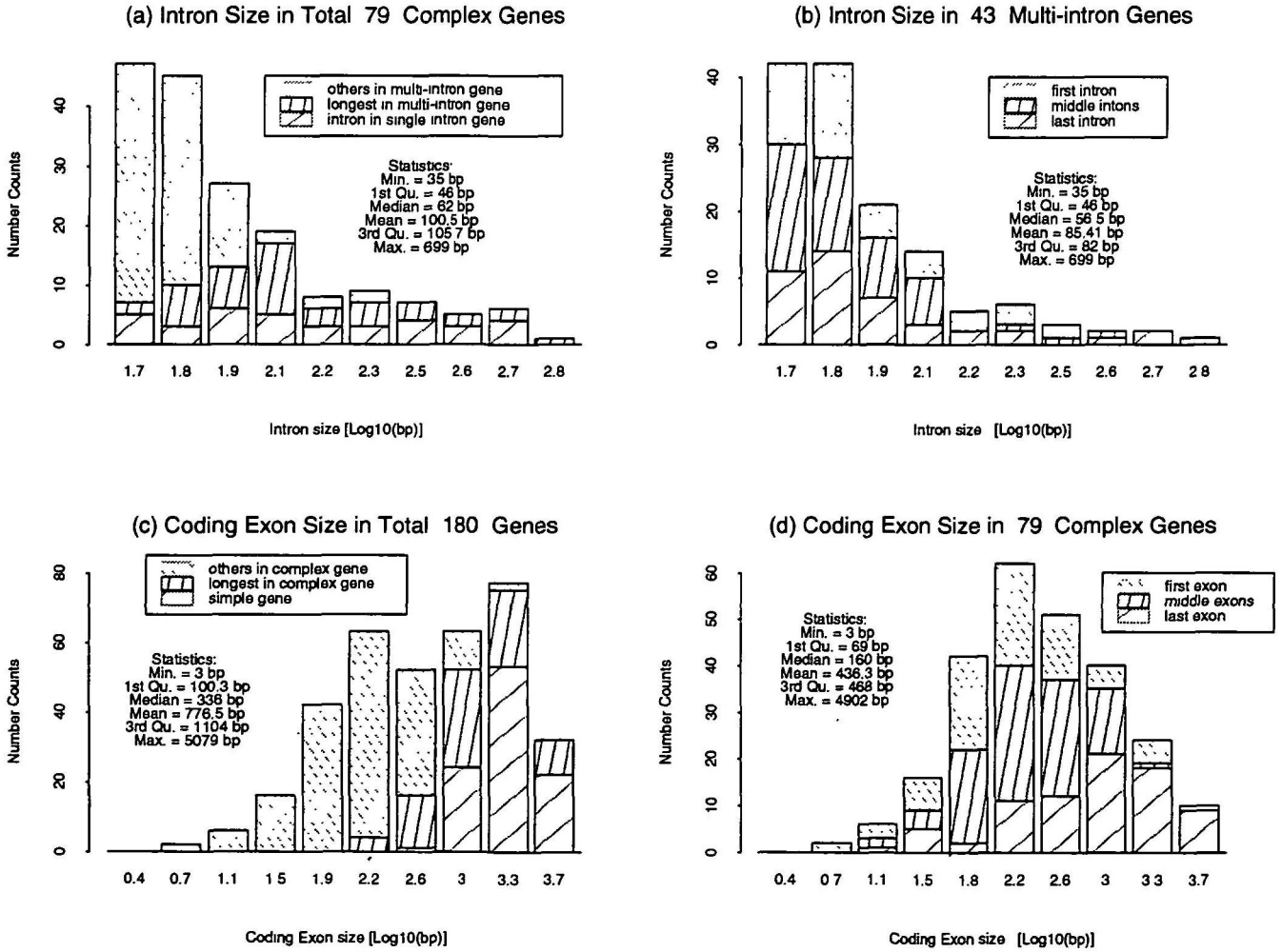

Figure 4. Size distributions of introns and coding exons: (a) Intron size distribution in total 79 complex genes; (b) Intron size distribution in the 43 multi-introns genes; (c) Coding exon size distribution in total of 180 genes; (d) Coding exon size distribution in 79 complex genes.

their distribution centers at 3.2 (1585 bp) which is the same for the distribution of the longest exons from each gene (Fig. 4c). It has been proposed that the coding sequences are selected from random ORF's by a minimum length cut-off constraint [42], and the long mRNA could only exist by splicing out the intervening sequences (containing clusters of in-frame stop codons) between random ORF's [43]. The tail distribution (> 100 bp) is clearly nonrandom where the coding sequences are likely to be found.

CDS size, distribution. In Figure 6, we plot the CDS size distribution. 210 sequences are divided into two groups: (1) 101 simple genes, (2) 79 spliced complex genes and (3) 30 cDNA

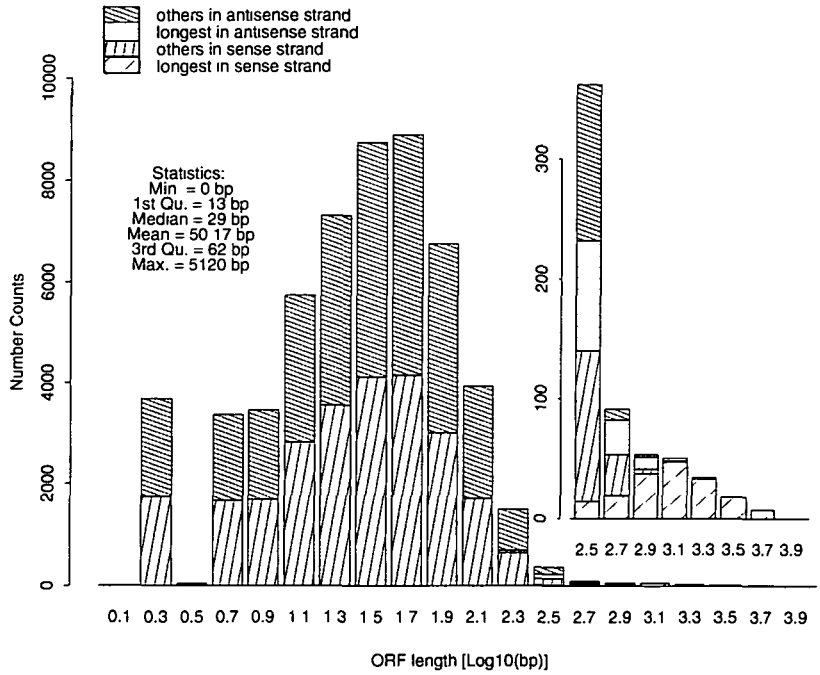

Figure 5. Open reading frame length distribution in 180 genomic sequences.

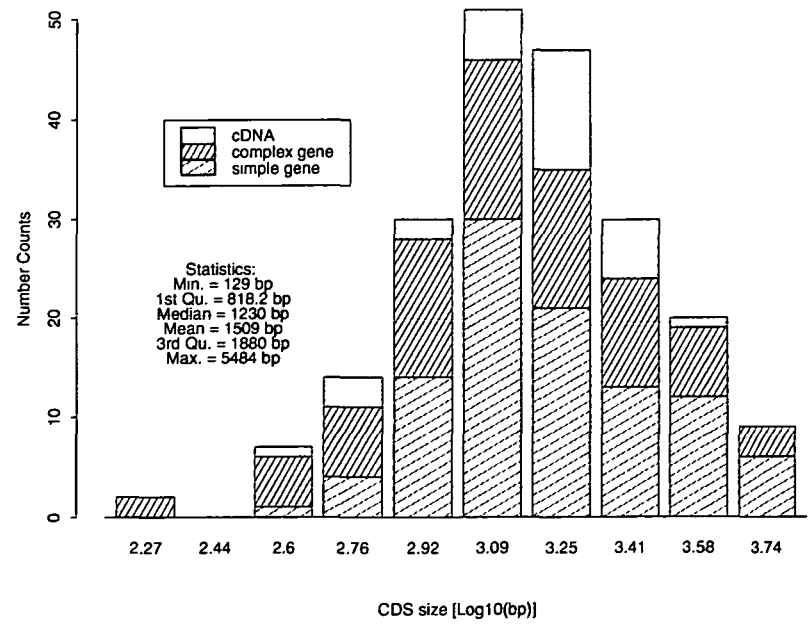

Figure 6. Fission yeast CDS size distribution in 210 genes.

Table 1. Fission yeast genes[1]

act1.Y00447.s	act2.M81068.c	ade1.X06601.s	ade2.M98805.c	ade4.X72293.s
ade6.X14488.s	adh1 J01341.s	adk1.X70363.s	arf1.L09551 r	arg7.X63262.s
atp.X59947.c	bip1.X64416.s	bws1.M27075.c	byr1.X07445.s	byr2^{ste8}.M74293.s
byr3.S45038.s	cal1.M98799.r	cam1.M16475.c	cap L16577.r	car1^{sod1}.Z14035.s
ccr1.X64702 r	cdc10 X02175.s	cdc13.X12557.s	cdc14 X72911.s	cdc16.X71605.c
cdc17.X05107.c	cdc18.L16793 s	cdc2.M12912 c	cdc21 X58824.s	cdc22.X65116.c
cdc25.M13158.s	cdc27.S84911.c	cdc42.L25677.c	cdc8.L04126.c	cdr1^{nim1}.X57549.s
cgs1.S64905.c	cgs2.S64907.c	chk1.L13742.c	cig1.M68881.s	cig2.X70046.r
crm1.X15482.s	csk1.S59896.c	cta3.J05634.s	cut1.M36179.c	cut2.M57750.s
cut5.D16627.c	cut7.X57513.c	cwg2.Z12155.s	cyc1.J01318.s	cyc19.???.c
cyr1.M26699 s	dbp2.X52648.c	dfr1.L13703.r	dis2.M27068.c	dis3.M74094.c
dsk1.D13447.s	erg9.L06071.r	esc1.X69389.r	fba.D17415 s	fbp1.J03213.s
fib.X69930.r	frp1 L07749.s	gap1^{src1}.D10457.s	gar1.Z19576.c	gpa1.M26699.s
gpa2.D13366 c	gpd1.X56162.s	gsa1.M85179.s	gtp1.L00671.c	his3.L19523.s
hmt1.Z14055.r	hsf.M94683 r	ilv1.L11293.s	kin1.M64999.s	leu1.M36910.s
mam2.X61672.s	map3.D10933.s	mat1.U02280.c	mcs2.S59895.c	mei2.X07180.s
mei3.X05142.s	mfm1.S87829.c	mfm2.S87831.c	mik1.M60834.c	mts2.Z29366.r
nda2.K02841.c	nda3.M10347.c	nmt1.J05493 r	ntf1.L25912.s	nuc1.M37411.s
nuc2.X07693.c	obr1.X73558.s	orf1 D14060.r	orf2.D14061.r	orf3.D14062.r
orf4.D14064.r	p25.D13038 s	pab1.M64603 r	pac1.X54998.s	pap1.X57078.s
pck1.D14337.c	pck2.D14338.s	pcn.X54857.c	pgm.X75385.r	pho1.M11857.r
pho2.X62722.s	pho4.X56939.s	php2.M63639.s	php3.X75072.r	pim1.M73528.c
pim2.???.c	pkc1.L07637.s	pma1.J03498.s	pma2.M60471.s	pmd1.D10695.s
pol1.X58299.c	pol3^{pold}.X59278 c	ppa1.M58518.s	ppa2.M58519.c	ppag.X54301.s
ppe1.D13712.s	ppi1.X53223.s	prh1.D13249.c	pro15.L25592.c	prp2a.Z18925.s
prp4.L10739.c	pts1.D13094.s	puc1.X74451.s	pyp1.M63257.r	pyp2.S51320.s
pyp3.X69994.c	rad1.M38132.c[2]	rad13.X66795.r	rad16.X71595.c	rad21.M96437.c
rad22.X72220.s	rad26 X76558.c	rad3.X63544 s	rad4.X62676.c	rad51.D13805.s
rad8.X74615.s	rad9.X64648.c	ral2.M30827.s	ran1.X04728.s	ras.X03771.c
rec7 M85297 s	rec8.M85298.s	res1^{sct1}.X68789.c[3]	rev.L10324.s	rhp3^{rad15}.X64583.s
rhp51.Z22691.s	rhp6 X53252.c	rpa1.M33137.c	rpa2^{rpl40}.M33138.c	rpa3.M33139.c
rpa4.M33142.c	rpb1.X56564.c	rpb2 D13337.c	rpb6.L00597.c	rpk37.X05036.s
rpk5.X51659.s	rpkd4.X16392.s	rpl2.X73146.s	rpl29.X57207.s	rpl3.X57734.s
rpl7a.X53575.c	rpl7b.X54983.c	rps13.X67030.c	rps6.M36382.s	rtp1.M38526.s
ryh1 X52475 c	sar1.M95797.r	sct1 L09270.r	sds21.M27069.c	sds22.M57495.s
sep.D14063.r	sod2.Z11736.c	spi1.M73527.c	spk1.X57334.s	spp31.???.c
spt.D14063.r	srp54 M55518.c	st11p.M95798.r	ste11.Z11156.s	ste4.X61924 c
ste6.X53254 s	sts1.X63549.s	suc1.M16032.c	suc22.X65115.s	swi10.X61926.c
swi4.X61306.c	swi6.X71783 s	sxa1.D10198.s	sxa2.D10199 s	tfd1.X53415.c
tmp.X65868 r	tms1.X74422.s	top1.X06201 c	top2.X04326.s	tpi1.M14432.s
trr1 ??? c	tub1.K02842.s	tug1^{gtb1}.M63447.c	u2af.L22577.s	ura3.X65114.r
ura4.M36504 s	vma1 X68580.c	vma2.X69638.c	wee1.M16508.s	wis1.X62631.s
ypt1.X52099.c	ypt2.X52469.s	ypt3^{gtpb}.X52100 c	ypt5.Z22220.c	

214 Fission yeast genes: 101 simple genes (*s*), 79 complex genes (*c*) and 30 cDNAs (r). The underlined were obtained privately before publications (*csk*1 and *mcs*2 were from E.Molz, *chk*1 from N.Walworth, *cyc*19 from D.Casso, *pim*2 from T.Matsumoto, *spp*31 from D.Frendewey and *trr*1 from T.Connolly). [1]: *rad*1 may contain two upstream introns (see text); [2]: *res1*^{sct1} contains an 127 bp intron (personal communication with M.Caligiuri).

sequences. They all have similar distributions, with a combined median of 1230 bp, although the mean size is only 1509 bp. Despite of having short introns in the genes, the gene products of the fission yeast have a typical size of 400−500 amino acids, which is longer than the average reported for Vertebrates (300 amino acids) [5].

Codon usage and 6-tuple frequency

Codon usage. We have prepared the codon usage table (Table 2) from 105,662 codons found in our data base of 210 genes. The most abundant was the GAA, codon of Glutamine (4.3%) and the least abundant, aside from the stop codons, was the CGG of Arginine (.3%). The most abundant amino acid was Leucine (9.5%) and the least abundant was Tryptophan (1.0%). Among the three stop codons, Ochre (UAA) was the most frequent and opal (UGA) was the least. It has been reported that codon usage depends on the expression level of the genes, and it is most biased in highly expressed genes [44]. Highly expressed genes almost exclusively terminate with the ochre codon [45]. It is known that codon usage is also related to G+C content (see [46] for a recent

Table 2. Codon usage in 214 fission yeast genes (total number of codons: 105,662)

Phe(F)	UUU	3097	Ser(S)	UCU	3167	Tyr(Y)	UAU	2281	Cys(C)	UGU	941
	UUC	1438		UCC	1273		UAC	1336		UGC	632
Leu(L)	UUA	2529		UCA	1693	Ter(*)	UAA	131	Ter(*)	UGA	32
	UUG	2537		UCG	745	Ter(*)	UAG	47	Trp(W)	UGG	1007
Leu(L)	CUU	2754	Pro(P)	CCU	2379	His(H)	CAU	1698	Arg(R)	CGU	1950
	CUC	793		CCC	946		CAC	701		CGC	683
	CUA	771		CCA	1239	Gln(Q)	CAA	2944		CGA	772
	CUG	646		CCG	422		CAG	1095		CGG	284
Ile(I)	AUU	3899	Thr(T)	ACU	2479	Asn(N)	AAU	3361	Ser(S)	AGU	1389
	AUC	1516		ACC	1288		AAC	1960		AGC	926
	AUA	1117		ACA	1401	Lys(K)	AAA	3852	Arg(R)	AGA	1132
Met(M)	AUG	2379		ACG	629		AAG	2735		AGG	456
Val(V)	GUU	3257	Ala(A)	GCU	3356	Asp(D)	GAU	4084	Gly(G)	GGU	2909
	GUC	1356		GCC	1421		GAC	1734		GGC	1031
	GUA	1234		GCA	1555	Glu(E)	GAA	4590		GGA	1698
	GUG	806		GCG	476		GAG	2307		GGG	366

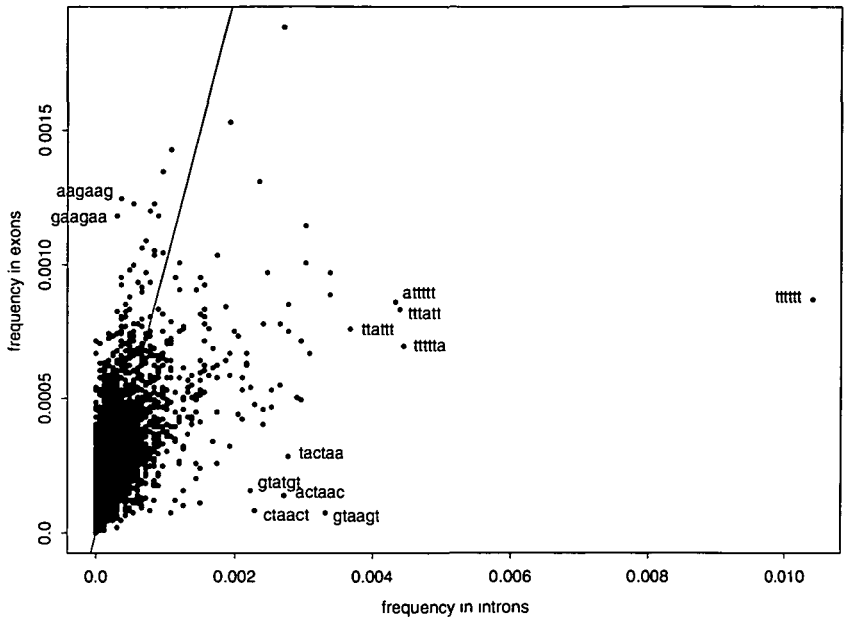

Figure 7. Distribution of 6-tuples in exons vs. introns (in 79 complex genes).

review), the G+C content for the fission yeast is 37% (the overall composition from our data set is A, 30%; C, 19%; G, 18% and T, 33%), which is about the same as (39–40%) the budding yeast. Codon usage in the absence of any selection will reflect the underlying mutational biases of the genome. Any selection pressures associated with translational differences among synonyms would add to this apparent mutational bias. The G+C content at third codon positions was found to be 34% (the composition at third codon positions was A, 25%; C,18%; G, 16% and U, 41%), the noticeable difference from the genomic composition may be caused by a sample bias. While for the budding yeast, the G+C content at third codon positions in lowly expressed genes was found to be 38%, which is very close to the genomic value [44,45]. We should point out that the composition for the fission yeast in exons was found to be A, 29%; C, 20%, G, 21% and T, 30%.

6-tuple frequency. It has been recognized for some time that 6-tuple frequency-based methods have the best discriminant power in exon/intron recognition [47]. This information can be also very useful for various primer selection methods and restriction endonuclease selection methods. We have computed the 6-tuple frequencies separately for exons and introns in complex gene sequences, [24] (there were 16,614 of intron 6-tuples and 109,112 of exon 6-tuples). The intron 6-tuples had a larger information value (The information was defined as $I[p] = \Sigma p \log(p)$, intron had −7.5 and exon had −8.1). There were 9 6-tuples (out of 4096 possible ones) absent in exons, they were CCCGCA, CGGCCC, ACCCGC, GCGGGC, CGCGCG, GGGGCG, GCGCGG, GGCGGG and CCGCGT and they all contained the rare dinucleotide CG (even though the fission yeast is not considered as a methylation regulated organism). The top frequency 6-tuples in exons were all rich in A, the ones that had frequency larger than .12% were AAAAAA, GAAAAA, TGAAAA, AAGAAA, AAAGAA, AAAAAG, AAGAAG, AAAAAT and AAAATT. In introns, there were 948 6-tuples absent, although they were less significant because of the limited data set. The top frequency 6-tuples in introns were highly significant, they consist of T stretches and splicing signals (Fig. 7). The difference of the 6-tuple frequencies between exons and introns can be best seen in Fig. 7, where they are plotted against each other for each 6-tuple [24]. From this plot, it is clear that the donor signals GTAAGT, GTATGT; branch site signals ACTAAC, TACTAA, CTAACT, in introns and some unknown signals AAGAAG, GAAGAA in exons are the 6-tuples that have the best discriminant power.

INTRON.PLOT: an interactive intron finding program

The main features of our INTRON.PLOT computer program are the following:

(1) We defined an intron scoring function $S(s)$ of a 6-tuple s to be

$$S(s) = \frac{f_i(s) - f_e(s)}{f_i(s) + f_e(s)}$$

where $f_i(s)$ and $f_e(s)$ be the frequencies of the 6-tuple s in introns and exons, respectively. This function varies between −1 and 1. It has the property that if $S(s) > 0$ implies s is more likely to occur in introns.

(2) We defined a 5′ splice site scoring function $S5(s)$ of a potential donor site s(with 9 nucleotides and G,T at positions 4,5) as

$$S5(s) = \sum_{i=1}^{9} \log\left(f5_i(s_i) + \frac{1}{126}\right) - 9\log p_0,$$

which is essentially the relative information ($f5_i$(s) is the donor profile represented by the positional mononucleotide frequency

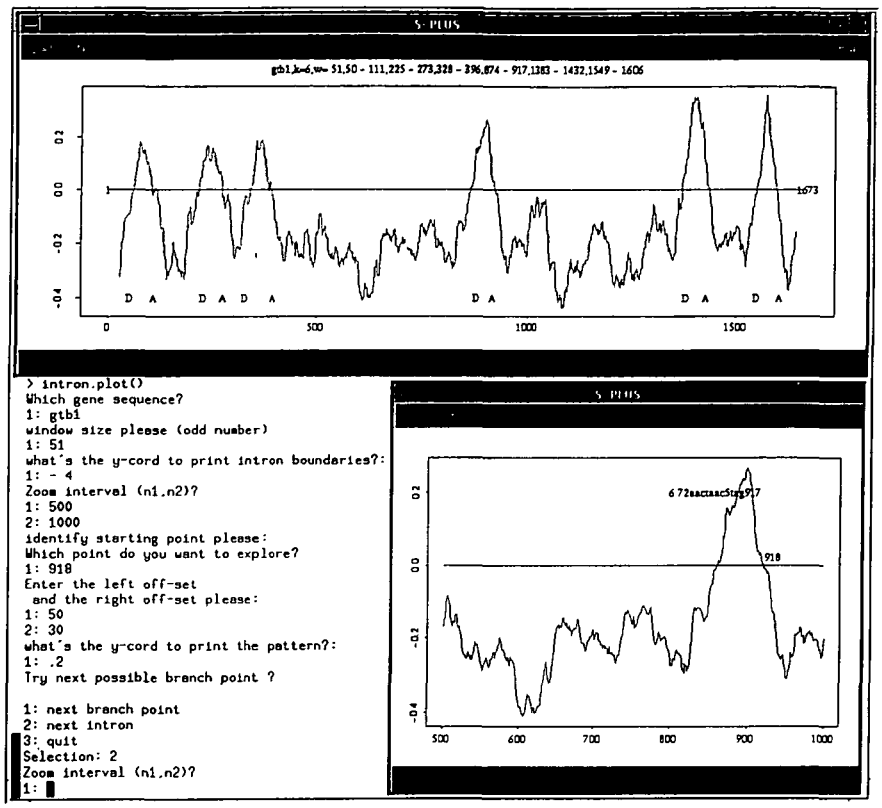

Figure 8. A screen-dump of an INTRON.PLOT (written in S-PLUS scripts) run on a DECstation 5000.

matrix (see the donor profile subsection above) or the positional dinucleotide frequency array ([48]) and p_0 is the corresponding uniform background frequency constant).

(3) We similarly defined a 3′ splice site scoring function $S3(s)$ of a joint potential branch site + variable distance + acceptor motif s(where a potential branch site is a 7-mer with an A at position 6; a variable distance is any interval of length between two parameters (we chose [3,25]; and a potential acceptor is the first 3-mer downstream of the branchpoint without a G at position 1 but with A,G at 2,3) as the relative branch site information:

$$S3(s) = \sum_{i=1}^{7} \log\left(f3_i + \frac{1}{126}\right) - 7\log p_0.$$

In a recent report [49], the scanning and competition between AGs in 3′ splice site selection of mammalian introns was discussed. The data indicated that proximity to the branch point is a major determinant of competition, although steric effects render an AG less competitive in close proximity ($\sim$12 nt). In addition, the nucleotide preceding the AG has a striking influence, the order of competitiveness is $CAG \simeq$ UAG > AAG > GAG in mammalian introns.

(4) We also defined a start site scoring function $S(s)$ of a potential start site s(a 9-mer with A,T,G at positions 6,7,8) as the relative information using the translational start profile.

Let E_{min}, be a parameter specifying the minimum size of the longest coding exon in a transcript (we chose it to be 100 bp) and I_{min} be another parameter specifying the minimum size of an intron (we chose 30 bp), we summarize briefly the main steps in prediction:

(1) Starting with the strand that has the longest ORF and choosing an appropriate window size WS (depending the ORF size distribution on the strand, we chose between 31 − 71 bp, 51 bp worked for most of the fission yeast genes. The lessen is that one does not want the window size too large to miss the signals or too small to increase noises), one plots the average 6-tuple scores in each moving window as the sequence is scanned.

(2) Picking the longest exon by judging from the overlap of the longest ORF and the large negative score region (if a long ORF interrupted by a single stop overlaps with a large negative S score region, there is probably a sequencing error; on the other hand, if a long ORF overlaps with a large negative S score region interrupted by a short intron-like positive score region without good splicing signals, the intron-like region may be just a false-positive). Together with the E_{min} constraint, one tries to find a matching pair (within I_{min}) of 5′ and 3′ splice sites of highest scores (measured by $S5$ and $S3$) in an upstream intron peak region. If one cannot find such a pair with satisfying scores (after setting a threshold value), one may assign the first in-frame ATG to be the translational start (if there is another inframe ATG downstream within a threshold distance, one should take the one that has the higher SS score); otherwise, one continues to the next upstream exon and repeats.

(3) One then starts from the downstream intron peak and works similarly towards the 3′ end, until the gene cannot be extended by finding a high score matching pair of splice sites and thus, terminate the coding region at the first stop codon.

This strategy is somewhat similar to SORFIND [17]. One should optimize various threshold parameters for a specific organism. In Fig. 8, we present a screen-dump from a typical run of INTRON.PLOT (ORF plot was not turned on) for *gtb1* which has 6 introns (their positions are indicated in the top window). The bottom window is a zoom-in of intron IV region where a 3′ splice site motif (represented by *S*3-score + branch site + variable distance + 3-mer acceptor + position of the 3′ end of the intron). A 5′ splice site can be plotted in a similar fashion.

In applying INTRON.PLOT to the 7 new genes: the underlined genes in Table 1. (*chk1, csk1, mcs2* were analyzed by INTRON.PLOT before submission), 18 out of the total 19 introns were identified precisely on the first attempt with a window size of 51 bp, and confirmed by the cDNA sequences. The only one missing in the initial attempt was a short (39 bp) inframe intron, which was then detected by using a window size of 31 bp. Some of the intron splicings (in *chk1* gene) were only detected experimentally after certain artifacts of the cDNA clone library was corrected. With the simple rule stated above, it was surprising that all the predicted translational initiators and terminators also appeared to be correct, except one start codon ATG which was disrupted by an intron after the first nucleotide A (in *trr1* gene), (an upstream inframe ATG was initially predicted). This tool is also very useful for detecting annotation errors in a database. For example, *rad1* was annotated as an intronless gene (nucleotide position 1026..1610) in GenBank. We found there are most likely two upstream introns missed, the gene position should be 470..605,732..972,1016..1610 [50].

ACKNOWLEDGEMENTS

We thank S.Cozza for computer system support. We thank D.Frendewey, T.Matsumoto, D.Casso, T.Connolly, N.Walworth and E.Molz, for providing their new sequences before publications. We thank M.Eggert, for providing us the RTIDE codes. We also thank the anonymous referee for pointing out some confusions in our initial manuscript. Financial support was kindly provided by NIH grant 1K01 HG00010-01 to M.Q.Z. and by NIH grant 1R01 HG00203-01A1 and DOE grant DE-FG02-91ER61190 to T.G.M.

REFERENCES

1. Chow,L.T., Gelinas,R.E., Broker,T.R. and Roberts,R.J. (1977) *Cell* **12**, 1–8; Berget,S.M., Moore,C. and Sharp,P.A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3171–3175.
2. Moore,M.J., Query,C.C. and Sharp,P.A. (1993) In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World*, Cold Spring Harbor Laboratory Press, New York, pp. 303–357; Weiner,A.M. (1993) *Cell* **72**, 161–164; Brown,J.D., Plumpton,M. and Beggs,J.D. (1993) *Antonie Van Leeuwenhoek* **62**, 35–46.
3. Moore,M.J. and Sharp,P.A. (1993) *Nature* **365**, 364–368.
4. Steitz,J.A. (1992) *Science* **257**, 888–889; Wassarman,D.A. and Steitz,J.A. (1992) *Science* **257**, 1918–1922.
5. Smith,M.W. (1988) *J. Mol. Evol.* **27**, 45–55.
6. Newman,A.J. and Norman,C. (1992) *Cell* **68**, 743–754; Reich,C.I., VanHoy,R.W., Porter,G.L. and Wise,J.A. (1992) *Cell* **69**, 1159–1169.
7. Gatermann,K.B., Hoffmann,A., Rosenberg,G.H. and Käufer,N.F. (1989) *Mol. Cell. Biol.* **9**, 1526–1535.
8. Parker,R. and Patterson,B. (1987) In Inouye,M. and Dudock,B.S. (eds), *Molecular Biology of RNA: New Perspective*. Academic Press, New York, pp. 133–149.
9. Prabhala,G., Rosenberg,G.H. and Käufer,N.F. (1992) *Yeast* **8**, 171–182.
10. Gallwitz,D., Halfter,H. and Mertins,P. (1987) In Kinghorn,J.A. (ed.), *Gene Structure in Eukaryotic Microbes*. IRL Press, Oxford, Washington DC, pp.27–40; Teem,J.L., Abovich,N., Käufer,N.F., Schwindinger,W.F., Warner,J.R., Levy,A., Woolford,J., Leer,R.J., van Raamsdonk-Duin,M.M.C., Mager,W.H., Planta,R.J., Schultz,L., Friesen,J.D., Fried,H.M. and Rosbash,M. (1984) *Nucleic Acids Res.* **12**, 8295–8315; Woolford,J.L. (1989) *Yeast* **5**, 439–457.
11. Fields,C. (1990) *Nucleic Acids Res.* **18**, 1509–1512.
12. Mizukami,T., Chang,W.I., Garkavtsev,I., Kaplan,N., Lombardi,D., Matsumoto,T., Niwa,O., Kounosu,A., Yanagida,M., Marr,T.G. and Beach,D. (1993) *Cell* **73**, 121–132.
13. Mount,S.M., Burks,C., Hertz,G., Stormo,G.D., White,O. and Fields,C. (1992) *Nucleic Acids Res.* **20**, 4255–4262.
14. Fields,C. and Soderlund,C. (1990) *CABIOS* **6**, 263–270.
15. Uberbacher,E.C. and Mural,R.J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11261–11265.
16. Guigo,R., Knudsen,S., Drake,N. and Smith,T. (1992) *J. Mol. Biol.* **226**, 141–157.
17. Hutchinson,G.B. and Hayden,M.R. (1992) *Nucleic Acids Res.* **20**, 3453–3462.
18. Snyder,E.E. and Stormo,G.D. (1993) *Nucleic Acids Res.* **21**, 607–613.
19. Cigan,A.M. and Donahue,T.F. (1987) *Gene* **59**, 1–18.
20. Stormo,G.D. (1990) *Methods Enzymol.* **183**, 211–221; (1988) *Annu. Rev. Biophys. Biochem.* **17**, 241–263.
21. S-PLUS is a rich graphical data analysis system and object-oriented programming language developed at Statistical Sciences, its prototype — the S language (developed at AT&T) may be found in Becker,R.A., Chambers,J.M. and Wilks,A.R. (1988) *The New S Language*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California.
22. Galas,D.J., Waterman,M.S. and Eggert,M. (1985) *J. Mol. Biol.* **186**, 117–128.
23. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) *J. Mol. Biol.* **188**, 415–431.
24. Claverie,J.-M., Sativaget,I. and Bougueleret,L. (1990) *Methods Enzymol.* **183**, 237–252.
25. Kozak,M. (1986) *Cell* **44**, 283–292.
26. Russell,P. and Nurse,P. (1986) *Cell* **45**, 781–782.
27. Kozak,M. (1992) *Crit. Rev. Biochem. Mol. Biol.* **27**, 385–402; (1991) *J. Biol. Chem.* **266**, 19867–19870.
28. Mount,S.M. (1982) *Nucleic Acids Res.* **10**, 459–472.
29. Iida,Y. (1990) *J. Theor. Biol.* **145**, 523–533.
30. Sakuraba,H., Eng,C.M., Desnick,R.J. and Bishop,D.F. (1992) *Genomics* **12**, 643–650.
31. Jacob,M. and Gallinaro,H. (1989) *Nucleic Acids Res.* **17**, 2159–2180.
32. Mount,S.M., Pettersson,I., Hinterberger,M., Karmes,A. and Steltz,J.A. (1983) *Cell* **33**, 509–519; Padgett,R.A., Grabowski,P.J., Kotiarska,M.M., Seiler,S.R. and Sharp,P.A. (1986) *Annu. Rev. Biochem.* **55**, 1119–1150.
33. Porter,G., Brennwald,P. and Wise,J.A. (1990) *Mol. Cell. Biol.* **10**, 2874–2881.
34. Woolford,J.L. (1989) *Yeast* **5**, 439–457.
35. Brown,J.W. (1986) *Nucleic Acids Res.* **14**, 9549–9558; Brown,J.W., Feix,G. and Frendeway,D. (1986) *EMBO J.* **5**, 2749–2758.
36. Nakata,K., Kanehisa,M. and DeLisi,C. (1985) *Nucleic Acids Res.* **13**, 5327–5340; Shapiro,M.B. and Senapathy,P. (1987) *Nucleic Acids Res.* **15**, 7155–7174.
37. Green,M.R. (1986) *Annu. Rev. Genet.* **20**, 671–708.
38. Brennwald,P., Porter,G. and Wise,J.A. (1988) *Mol. Cell. Biol.* **8**, 5575–5580.
39. Erös,P. and Rényi,A. (1970) *J. Analyse Math.* **23**, 103–111.
40. Cleveland,W.S. (1979) *J. Am. Stat. Assoc.* **74**, 829–836; Chambers,J.M., Cleveland,W.S., Kleiner,B. and Tukey,P.A. (1983) *Graphics Methods for Data Analysis*. Wadsworth, Belmont, California. And *S-PLUS Reference Manual* 3.0 (1991), Statistical Sciences, Inc., Seattle, Washington.
41. Hawkins,J.D. (1988) *Nucleic Acids Res.* **16**, 9893–9908.
42. Höglund,M, Säll,T. and Röhme,D. (1990) *J. Mol. Evol.* **30**, 104–108.
43. Senapathy,P. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2133–2137.
44. Bennetzen,J.L. and Hall,B.D. (1982) *J. Biol. Chem.* **257**, 3026–3031; Sharp,P.M., Tuohy,T.M.F. and Mosurski,K.P. (1986) *Nucleic Acids Res.* **14**, 5125–5143.
45. Sharp,P.M. and Cowe,E. (1991) *Yeast* **7**, 657–678.
46. Osawa,S., Jukes,T.H., Watanabe,K. and Muto,A. (1992) *Microbiol. Rev.* **56**, 229–264.
47. Ficket,J.W. and Tung,C.S. (1992) *Nucleic Acids Res.* **20**, 6441–6450.
48. Zhang,M.Q. and Marr,T.G. (1993) *CABIOS* **9**, 499–509.
49. Smith,C.W.J., Chu,T.T. and Nadal-Ginard,B. (1993) *Mol. Cell. Biol.* **13**, 4939–4952.
50. P.Young, private communication.