Nucleic Acids Research, 1994, Vol. 22, No. 7 1247–1256

Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix anaylsis of transcriptional control regions

Angelo Pavesi, Franco Conterio, Angelo Bolchi[1], Giorgio Dieci[1] and Simone Ottonello[1,*]
Department of Evolutionary Biology and [1]Institute of Biochemical Sciences, University of Parma, I-43100 Parma, Italy

Received November 19, 1993; Revised and Accepted February 28, 1994

ABSTRACT

A linear method for the search of eukaryotic nuclear tRNA genes in DNA databases is described. Based on a modified version of the general weight matrix procedure, our algorithm relies on the recognition of two intragenic control regions known as A and B boxes, a transcription termination signal, and on the evaluation of the spacing between these elements. The scanning of the eukaryotic nuclear DNA database using this search algorithm correctly identified 933 of the 940 known tRNA genes (0.74% of false negatives). Thirty new potential tRNA genes were identified, and the transcriptional activity of two of them was directly verified by *in vitro* transcription. The total false positive rate of the algorithm was 0.014%. Structurally unusual tRNA genes, like those coding for selenocysteine tRNAs, could also be recognized using a set of rules concerning their specific properties, and one human gene coding for such tRNA was identified. Some of the newly identified tRNA genes were found in rather uncommon genomic positions: 2 in centromeric regions and 3 within introns. Furthermore, the presence of extragenically located B boxes in tRNA genes from various organisms could be detected through a specific subroutine of the standard search program.

INTRODUCTION

A significant issue in the statistical analysis of genomic DNA sequences regards the identification of individual classes of transcription units interspersed within uncharacterized or otherwise unrelated DNA sequences (1,2). This also pertains to the recognition of the relatively small transcription units coding for eukaryotic tRNAs. Indeed, tRNA genes have been found in particular positions within the nuclear genome such as in centromeric regions (3,4) and in close proximity to genes transcribed by RNA polymerase II (5,6).

Various computer methods for the identification of tRNA coding genes have previously been described. Most of them deal primarily with secondary structure prediction and are thus based on RNA folding, carried out by means of either minimum free energy calculations (7) or base pairing rules (8,9). The occurrence at particular positions of bases that are common to all or most tRNAs, the so called invariant or seminvariant bases, has also been exploited for this type of analysis (8–10). A more exhaustive prediction algorithm based on the detection of seven distinct patterns, mainly related to the cloverleaf model of tRNA structure, has recently been reported (11). An alternative, but possibly complementary, searching approach, allowing the screening of DNA sequence data in a secondary structure-independent fashion and thus capable of detecting also tRNA genes with some imperfect patterns, could exclusively rely on the recognition of correctly positioned transcriptional control elements. These elements, two conserved intragenic control regions, the A box (also known as 5′-ICR, 'GTGGCNNAGT---GGT--AGNGC') and the B box (also known as 3′-ICR, 'GGTT-CGANTCC'), and a transcriptional termination signal consisting of a stretch of 4 to 6 thymine residues placed downstream of the B box, are present in all tRNA genes (12,13). Given the partially degenerate character of the A and B box sequences, a simple matching with their consensus sequences does not provide enough sensitivity and can easily miss authentic tRNA genes. However, more resolving procedures of signal search, such as the weight matrix analysis of Staden (14), in combination with the search for a less degenerate signal, like the RNA polymerase III termination site (for which the only pattern to be weighted is its distance from the end of the gene), could considerably enhance the reliability of a search procedure based on the recognition of these transcriptional control elements.

Here, we present an algorithm for the search of nuclear eukaryotic tRNA genes that is based on the detection and on the evaluation of the correct positioning of the above described control elements. A multistep weight matrix procedure, which accurately defines the insertion–deletion pattern of the A box and also takes into account the heterogeneous spacing between transcriptional control elements, was developed. The computer search program derived from this algorithm correctly recognized 933 out of 940 known eukaryotic tRNA genes, including the

*To whom correspondence should be addressed

structurally unusual genes coding for selenocysteine tRNAs. Thirty new potential tRNA genes were identified, and the presence of extragenically located B boxes, a motif found only in *Dictyostelium discoideum* tRNA genes so far (15), was detected in a variety of organisms.

EXPERIMENTAL PROCEDURES

Description of the algorithm

A collection of 115 aligned ICRs, with appropriately positioned gaps at the level of A boxes, was used as an initial reference (12). To obtain a more comprehensive list of ICRs, fully sequenced tRNA genes were selected from DNA databases (EMBL 27 and GenBankK 67; 16,17) by standard retrieval software (PC/GENE release 6.6). Using previously reported nucleotide frequency tables (12), we then subjected this updated 'tDNA file' to weight matrix analysis to identify A and B boxes. Information concerning the position of gaps was included in this analysis to achieve an optimal alignment of A boxes. Redundant sequences and pseudogenes were excluded from this preliminary scan to generate a final training set of 231 distinct tRNA genes. The occurrence of individual nucleotides and gaps at various positions of the A and B boxes in our training set is reported in Table 1. For each tRNA gene, the nucleotide distance interposed between the A box and the B box, and between the B box and the transcriptional terminator was also determined. The absolute frequency distribution of both distances is reported in Table 2. On the basis of this set of data, A box and B box weight matrices (i.e. the natural logarithm of the relative base frequency values for each position) and two weight vectors (i.e. the natural logarithm of the relative frequency distribution value for individual classes of distances) were calculated (14). Individual scores, in the form of either the summation of natural logarithm values for each nucleotide position (A, B box weight matrices) or single natural logarithm values corresponding to the A–B box and B box-terminator distances, were assigned to each of the four distinct patterns that describe a tRNA gene: i) nucleotide composition of the A box; ii) nucleotide composition of the B box; iii) A–B box spacing; iv) B box-termination site distance. For example, the score value for a given A box or B box sequence was calculated by summing the natural logarithm values of the relative base frequency at each position as derived from weight matrices for either box. Similarly, scores for either A–B box or B box-terminator spacings correspond to the natural logarithm of the relative frequency distribution values (number of times a certain distance occurs/number of aligned sequences) of any given distance. Finally, a total score value which quantitatively evaluates all the above features was calculated, by summation, for each member of the training set.

As outlined in Figure 1, the search algorithm derived from the sequential examination of the above described patterns consists of four steps; each of these steps is characterized by an individual score value that is calculated as described above for the training set . In the initial step, potential B box sequences are identified by a first probability score. Boxes of eleven nucleotides having a score value greater than, or equal to, −14.14 (a cut-off value which includes all of the B boxes in our sample set) are accepted. The second and third step correspond, respectively, to the identification and position analysis of the A box at a distance comprised between 24 and 139 nucleotides upstream of the B box; this spacing allows for an intron of up to 113 nucleotides which is the maximum observed length for an eukaryotic nuclear tRNA gene (DMTGYC, EMBL entry name, 6). Both steps are defined by individual scores, which are then summed to the first score, to obtain an intermediate score value. The cut-off for the intermediate score was set at a value of −31.25 (corresponding to the $tRNA^{Ser}$ gene, ATPATY3), which includes 98.7% of the tRNA genes in our sample with the exclusion of three

Table 1.

Frequency table for the A box[1]

p[2]	7	8	9	10	11	12	13	14	15	16	17	17a	18	19	20	20a	20b	21	22	23	24	25
A	65	6	94	2	0	22	8	229	32	7	1	0	0	0	19	2	0	224	56	71	81	3
T	54	225	6	3	82	75	79	0	1	175	57	0	0	0	192	101	30	0	49	20	0	66
G	106	0	127	214	6	55	46	2	198	14	1	0	231	231	8	6	0	7	118	84	144	6
C	6	0	4	12	143	79	98	0	0	35	19	0	0	0	12	25	1	0	8	56	6	156
-[3]	0	0	0	0	0	0	0	0	0	0	153	231	0	0	0	97	200	0	0	0	0	0

Frequency table for the B box[1]

p[2]	52	53	54	55	56	57	58	59	60	61	62
A	22	1	16	0	0	53	231	104	19	0	7
T	6	0	215	231	1	2	0	81	156	1	27
G	194	230	0	0	0	176	0	27	1	1	6
C	9	0	0	0	230	0	0	19	55	229	191

[1]Absolute frequency data derived from the multiple alignment of A and B boxes of 231 nuclear tRNA genes were used to calculate the corresponding weight matrices (data not shown). To this end, individual values reported in the frequency tables were divided by the total number of aligned sequences and the natural logarithm of the resulting value was then calculated. In the case of bases that do not occur at a particular position in any of the 231 sample sequences, a value equal to the reciprocal of the total number of aligned sample sequences was used (14).
[2]P is the position of individual nucleotides within the cloverleaf model of secondary structure according to the standard numbering system (19).
[3](-) indicates an empty position.

selenocysteine tRNA ($tRNA^{Sec}$) genes. The fourth step checks for the presence of an RNA polymerase III transcription termination site containing at least four consecutive thymines; a spacing of 133 nucleotides between this element and the 3'-end of the B box, as present in a human tRNA gene (HSTRNS1, 18), is accepted as an upper limit for this distance. A third score, weighting the distance between the B box and this stretch of thymine residues, is then calculated and summed to the intermediate score to obtain the total score value (cut-off = −31.80). This final threshold, which corresponds to the score value of the $tRNA^{Ser}$ gene (ATPATY3), comprises 97.4% of the tRNA genes in our training set with the exclusion of all six $tRNA^{Sec}$ genes.

When the total score is equal to, or higher than, a critical cut-off value of −31.80, the sequence under examination is considered a potential tRNA gene. As a last step of the program, the boundaries of all sequences corresponding to potential tRNAs are inferred by moving 6 nucleotides upstream of the 5′ end of the A box and 11 nucleotides downstream of the 3′ end of the B box. Similarly, the anticodon region of each new putative tRNA gene is predicted at this stage of the analysis by shifting eight nucleotides downstream of the 3′ end of the A-box and considering the three subsequent nucleotides (19). Using this rule we could correctly predict the anticodon region for 99.25 % of the known tRNA genes that were recognized by our algorithm.

Table 2.

Frequency distribution of the distance between A and B boxes[1]

D(bp)[2]	N[4]
24−30	146
30−36	37
36−42	22
42−48	9
48−54	4
54−60	5
60−66	2
66−72	1
72−78	1
78−84	1
84−90	0
90−96	0
96−102	0
102−108	0
108−114	0
114−120	0
120−126	0
126−132	0
132−138	0
138−144	1

Frequency distribution of the distance between the B box and the transcription termination signal[1]

11−17	134
17−23	57
23−29	14
29−35	8
35−41	9
41−47	1
47−53	0
53−59	4
59−65	0
65−71	0
71−77	1
77−83	0
83−89	1
89−95	0
95−101	0
101−107	1
107−113	0
113−119	0
119−125	0
125−131	0
131−137	1

[1]Data are grouped as arbitrarily selected classes of distances.
[2]D is the distance between the last base of the A box and the first base of the B box.
[3]D is the distance between the last base of the B box and the first T residue of the transcription termination signal.
[4]N is the absolute frequency of individual classes of distances; weight vectors for A−B box and B box-terminator distances (i.e. the natural logarithms of the relative frequency distribution values for each class of distance) were calculated as in Table 1 (data not shown).

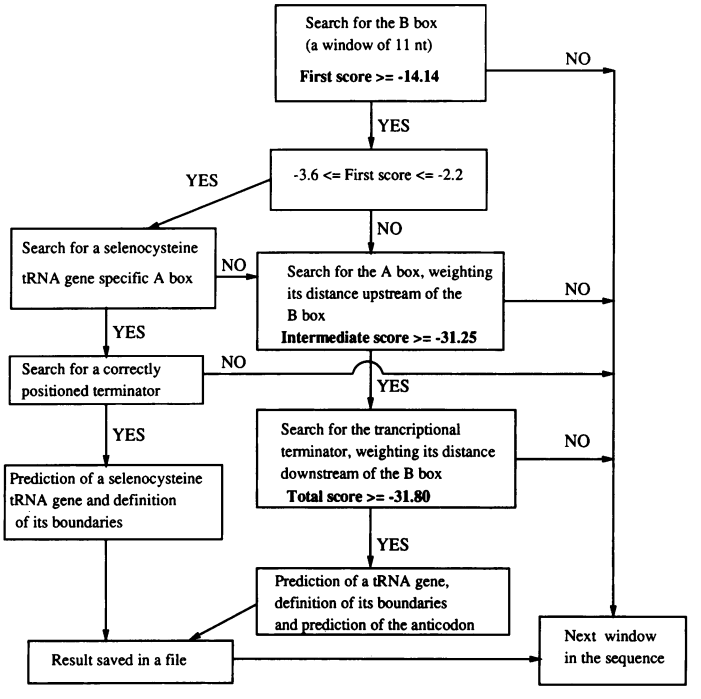

Figure 1. Schematic description of the algorithm. Individual steps of the algorithm and the corresponding threshold values are reported in separate boxes. If the score of the sequence under examination is not higher than the threshold at each step, the window is shifted by 1 nucleotide and the search is re-initiated on a new subsequence. The subroutine for selenocysteine tRNA ($tRNA^{Sec}$) gene recognition is also shown. When the score for the first step is comprised between −2.2 and −3.6, the search for a $tRNA^{Sec}$ gene-specific A box, at a distance of 36 nucleotides from the B box, is carried out; if a $tRNA^{Sec}$ gene A box is found, the algorithm proceeds with the search of a correct termination site (see Results for details). The identification by the algorithm of a known $tRNA^{Tyr}$ gene (ATPATT1) in a 210 bp sequence (position 71−155) can be exemplified as follows. In the first step, a potential B box sequence ('GGTTCGAATCC') with a score of −1.92 is found at position 134. Next, the search of an A box is carried out, and two potential A boxes ('TTGGTAGAGC--GGA--GGACT'; 'TTAGC-TCAGTT-GGT--AGAGC') are found 30 nt and 38 nt upstream of the B box, respectively. Since the corresponding intermediate scores for these two potential A boxes are −25.76 (sum of −1.92, −0.46, −23.38) and −14.97 (sum of −1.92, −2.35, −10.70), the one with the highest score value is selected by the program as the most likely. A correctly positioned termination site is then identified (31 nt downstream of the B box), and its individual score value (−3.36) is added to the intermediate score to obtain a total score value of −18.33, which is well above the final threshold value (−31.80). The predicted boundaries and anticodon of the tRNA gene identified in this way are identical to those previously reported for this particular $tRNA^{Tyr}$ gene.

The algorithm was applied to both strands of eukaryotic nuclear DNA sequences (EMBL Rel. 33 and GenBank Rel. 76) that were grouped into six different bulk files (fungi, plants, invertebrates, vertebrates, rodents and primates) using facilities included in the software package PC/GENE Rel. 6.7. All sequences with total score values above the threshold were compared with the entire set of known tRNA sequences by the FASTSCAN computer program (20). A refined alignment between each candidate sequence and the most similar tRNA sequence was obtained with the NALIGN program (21). Random permutations of all DNA sequences, maintaining their base composition, were generated by the GWBASIC randomization functions. The search of selenocysteine tRNA genes and of extragenically located B boxes was carried out using two specifically developed subroutines which were included into the standard program (see Figure 1 and Results). The tRNA gene search program was written in GWBASIC; it runs on standard MS-DOS compatible personal computers and the upper size limit of the sequence that can be handled is 30,000 nucleotides. Using a compiled version of the search program and a personal computer equipped with an Intel 486DX/33 microprocessor, it takes about two hours to scan both strands of a 318,444 bp sequence (a total of 32 entries, whose overall length corresponds to that of yeast chromosome III).

PCR amplification and *in vitro* transcription of new potential tRNA genes

DNA sequences corresponding to two newly identified putative tRNA genes [coding for $tRNA^{Leu(UAA)}$ and $tRNA^{Leu(CAA)}$] and a previously predicted gene coding for $tRNA^{Leu(UAG)}$ (11), were amplified under standard PCR conditions (22). A cosmid clone containing a distal portion of the *A.thaliana* enolase gene (23), plasmid pSS170 (24) and plasmid pEMBLYr25-MET25 (25) (kind gifts of D.Van der Straeten, M.Yanagida and Y.Surdin-Kerjan) were used as templates, with the following sets of oligonucleotide primers:

$tRNA^{Leu(UAA)}$ gene
(length of the amplified fragment: 253 bp)
(5′)CTTAATCAAGTGATGGTGAAGG(3′)
(5′)CCACAGCCTGCATATTTTTCC(3′)

$tRNA^{Leu(CAA)}$ gene
(length of the amplified fragment: 303 bp)
(5′)TTTTTGCAGTAGCATCAGCCA(3′)
(5′)CCCTAAAATGATAGCGAAGGA(3′)

$tRNA^{Leu(UAG)}$ gene
(length of the amplified fragment: 320 bp)
(5′)TTTGCCAACCACCACAGTTC(3′)
(5′)CTAGTTAGTAGATGATAGTTGAT(3′)

The 5′-ends of the six primers were positioned 98 bp, 131 bp and 127 bp upstream of the A box, and 44 bp, 66 bp and 57 bp downstream of the termination site, in the case of the $tRNA^{Leu(UAA)}$, $tRNA^{Leu(UAG)}$ and $tRNA^{Leu(CAA)}$ genes, respectively. Amplified DNA fragments were gel-purified and used to program *in vitro* transcription reactions carried out in the presence of a *Saccharomyces cerevisiae* nuclear extract, which has previously been shown to transcribe also a variety of plant tRNA genes (S.Ottonello, unpublished results; 26). Reaction mixtures contained 15 mM Tris−HCl (pH 7.9), 120 mM KCl, 6 mM $MgCl_2$, 10 % (v/v) glycerol, 600 μM each of ATP, GTP and CTP, 25 μM UTP, 5 μCi [α-^{32}P]UTP (NEN, 800 Ci/mmol), 1 μl of yeast nuclear extract and 100 fmol of each PCR-amplified DNA fragment in a final volume of 20 μl; they were incubated at 20°C for 10 minutes, then stopped, processed and analyzed as described (27). Reference tRNA genes from *S.cerevisiae*, carried on plasmids pJD137 ($tRNA^{Leu3}$ gene, 28), pJD110 ($tRNA^{Leu3}$ gene with a 21 bp insertion in the intervening sequence, 28) and pSArg ($tRNA^{Arg(AGG)}$ gene, 29), were individually transcribed (100 fmol each) under the same experimental conditions.

RESULTS

Test of the algorithm and detection of new putative tRNA genes

The histogram reported in Figure 2 shows the extent of variation of the total score values associated to the 231 tRNA genes of our sample set. All genes whose total score is close to, but below,

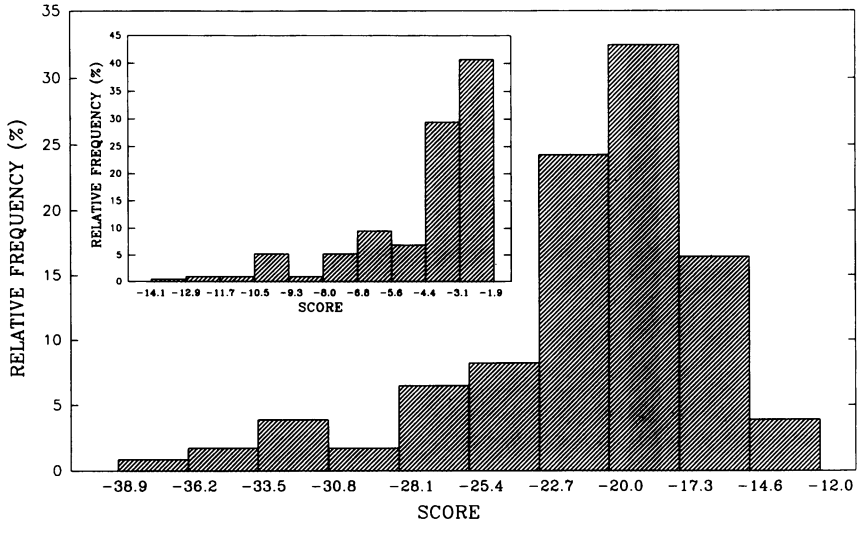

Figure 2. Frequency histogram of the total score values of a sample set containing 231 nuclear tRNA genes. The inset shows a similar frequency histogram for the first score values (B box sequences) only.

the threshold (−33.7 to −38.9, Figure 1) correspond to selenocysteine tRNA genes. As indicated by a detailed examination of these particular tRNA genes, their sub-threshold score values are specifically determined by deviation from consensus at the level of the A box region. Interestingly, B box weighting scores of $tRNA^{Sec}$ genes are all comprised in a narrow region of the B box distribution, ranging from −2.2 to −3.6 (Figure 2, inset). The six available sequences of $tRNA^{Sec}$ genes were thus aligned, and a clear consensus sequence ('GGTCYGG/TGGT') was localized in the distal part of the A box region, at a constant distance from the 5'-end of the B box. Based on these observations, a set of $tRNA^{Sec}$ specific rules was formulated: i) a B box score value ranging from −2.2 to −3.6; ii) a decanucleotide perfectly matching the consensus sequence 'GGTCYGG/TGGT' located 36 nucleotides upstream of the B box; iii) the presence of a correctly positioned transcriptional terminator. Using this set of rules, all of the known $tRNA^{Sec}$ gene sequences were correctly recognized and the corresponding subroutine was thus included into the standard search program (Figure 1).

At a cut-off value of −31.80, the search algorithm was then applied to the entire set of eukaryotic nuclear DNA sequences stored in genetic databases (EMBL Rel. 33 and GenBank Rel. 76) and its accuracy was tested following the procedure of Fichant and Burks (11). A summary of the general results obtained from this analysis is reported in Table 3. The occurrence of false-negative instances (i.e. unrecognized tRNA gene sequences) and of false-positive predictions (i.e. unrelated sequences predicted to be tRNA genes), as well as the number of false positives found in random sequence permutations (a number which should ideally correspond to zero) were determined for each group of sequences.

Out of 307 known tRNA genes in fungi, 305 were correctly recognized by the algorithm, yielding a false-negative rate of 0.65% (Table 3). The two genes which were missed are both part of tandemly organized transcription units whose primary transcript is a dimeric precursor that is post-transcriptionally processed into two distinct tRNA molecules. A $tRNA^{Arg}$ gene and a $tRNA^{Ser}$ gene are the upstream elements of two dimeric units that are present in SCTRNRYS and SPTGSUP, respectively (30,31). In both cases, the algorithm failed to detect the upstream tRNA gene due to the absence of a correctly positioned transcriptional terminator. An apparently false negative prediction was made in the case of SCLSC, an entry featured as a tRNA gene in the database, but corresponding to a mutagenized derivative of the SUP53 gene (32). Two critical base substitutions in the A and B box regions reduced the total score value of this sequence from −26 to −36. Seventeen previously unidentified sequences with total scores ranging from −15.4 to −28.6 were detected in fungi. Since they all displayed a high similarity with known tRNAs we considered these sequences as *bona fide* new tRNA genes (Table 4, but see Discussion). Only one false-positive, with a total score of −31.5 but exhibiting a low similarity (less than 70%, Table 3) with any known tRNA gene, was obtained from the screening of fungi sequences; the final rate of false-positive predictions for this taxonomic group was the lowest, and corresponded to 0.0006%.

A false negative rate of 0.93% was obtained from the scanning of plant sequences, and only one tRNA gene, out of 108 known, was missed by the algorithm in this taxonomic group (Table 3). This gene, TVTRNS (33), is a rarely occurring example of a tRNA gene that is transcribed as a single entity but lacks a stretch of at least four consecutive thymine residues at its 3'-end. Three

Table 3. General results of the database search

Taxonomic group	Number of nucleotides[1]	Known tRNA genes (number of nt)[2]	% false negative	Newly predicted tRNA genes[3]	False positives[4]	% false positive[5]	False positives in randomly permuted sequences[6]
Primates	21,164,080	6093	0.00	3(244)	17	0.0031	6
	22,291 entries		0 out of 78		(4 pseudogenes)		
Rodents	15,966,600	6104	0.00	0	149	0.0355	4
	14,224 entries		0 out of 83		(9 pseudogenes)		
Vertebrates	8,699,903	5850	1.30	0	33	0.0144	3
	6955 entries		1 out of 77				
Invertebrates	11,080,358	21,783	1.04	7(525)	28	0.0096	3
	9049 entries		3 out of 287		(2 pseudogenes)		
Plants	6,654,923	8510	0.93	3(243)	13	0.0083	0
	5353 entries		1 out of 108		(3 pseudogenes)		
Fungi	6,947,375	24,420	0.65	17(1448)	1	0.0006	1
	3708 entries		2 out of 307				
Total	70,513,245	72,760	0.74	30(2460)	241	0.0137	17
	65,180 entries		7 out of 940		(18 pseudogenes)		

[1]The indicated number of nucleotides (nt) corresponds to a single strand of DNA, but both strands were scanned by the program.
[2]The test set of 940 known tRNA genes includes 12 tRNA genes that were not annotated in the database, but that were later identified and reported in the literature. Their entry names are the following: SC1 (123−194), SCMCM3G (334−415), SCHYP2 (1362−1441), SCTIF51A (1319−1394), SCOASOAS (236−336), SCBAR1A (2713−2642; c), SCHAP1 (4992−4820; c), SCCYP1 (4891−4819; c), SCSTE6G (169−240), SCANBI (1995−2076), SPCEN114 (1774−1853), SPCEN114 (2663−2593; c). (c) indicates the location of a tRNA gene in the complementary strand.
[3]Values in parentheses indicate the number of nucleotides corresponding to new potential tRNA genes.
[4]Sequences classified as false positives, either share less than 70% similarity with known tRNAs or correspond to previously described tRNA pseudogenes.
[5]The false positive rate was calculated according to Fichant and Burks (11): values in column 2 and values parenthesized in column 4 were subtracted from those in column 1, the resulting value was then divided by average tRNA length and multiplied by 2 to obtain the number of tRNA-sized objects in the searched DNA. Finally, the number of false positives (column 5) was divided by the above calculated number of potential tRNAs in the negative set to obtain the false positive rate (column 6).
[6]One set of randomly permuted sequences was generated (see Experimental Procedures) and analyzed with the search program.

Table 4. New potential tRNA genes

Taxonomic group	Entry name	Accession number	Positions	Total score[1]	tRNA type[2]	Compared sequence	Similarity (%)
Fungi	NCGLA1	X67291	3470-3541 (f)[3]	-26.8	Arg(ACG)	DMRP	82
	PPURNA1	X12547	(c)[4]343-272 (f)	-26.2	Met(CAU)	MCTRFM	70
	PTGAL10	X68593	3354-3440	-27.1	Phe(GAA)	SCFT5	83
	SCAMDY	X56043(t)[5]	331-419 (f)	-28.0	Tyr(GUA)	YSTRYP	100
	SCCDC14A	M61194	705-793 (f)	-28.0	Tyr(GUA)	YSTRYP	1000
	SCCDC42	X51906	83-216 (f)	-19.6	Ile(UAU)	SCIP	99
	SCCEN11D01	X65124(t)	(c)3721-3608 (f)	-26.9	Leu(CAA)	SCTGL	100
	SCCLB5A	M91209(t)	385-456 (f)	-18.1	Cys(GCA)	SCTRNCYS	100
	SCGBEAA	M76739(t)	(c)2483-2412	-28.6	Lys(CUU)	SCK1	98
	SCGSH1	M87066	(c)3692-3620 (f)	-19.5	Arg(ACG)	SCRN15	97
	SCICL1G	X65554(t)	524-605 (f)	-16.9	Ser(UGA)	SCSM	98
	SCPP1A	M77175(t)	(c)2984-2913 (f)	-18.6	His(GUG)	SCTGHC	100
	SCTY109	X02417(t)	(c)6328-6227	-22.3	Trp(CCA)	SCTRWSIG	90
	SCTY4(1)[6]	X67284(t)	7447-7517	-19.3	Asp(GUC)	SCD	99
	SCTY4(2)[6]	X67284(t)	(c) 102-29	-14.8	Ala(AGC)	SCGUT2P	99
	SCUBPIII	M94917	(c)4245-4172 (f)	-20.6	Val(AAC)	SCTVTY3	100
	SPCEN114	X13761	(c)2103-2005	-27.1	Leu(CAA)	SCTGL	70
Plants	ATCPYLP	M81130	234-323 (f)	-16.7	Tyr(GUA)	NRTY8	83
	ATENGE	X58107	4380-4462 (f)	-23.0	Leu(UAA)	SCLEURNA	74
	GMSB1TUB	M21296	2758-2827	-18.3	Cys(GCA)	PCTRNA	83
Invertebrates	CEGPA1	M38249	2929-3012 (f)	-20.4	Leu(CAG)	DMTGL	77
	CEPGPAG	X65054	6762-6690 (f)	-17.1	Lys(CUU)	CETGKA	100
	CEUNC22_01	X15423	(c)1474-1545 (f)	-18.5	Pro(CGG)	CETGPX	93
	CEUNC33G	Z14148	9955-10027	-19.9	Arg(ACG)	CEARGTRNA	97
	EHDNA	M77091	(c) 508-432	-16.4	Glu(UUC)	EHTRGLUG	86
	EHRPTARQ	M55341	278-362 (f)	-15.5	Thr(UGU)	TATRTY1	85
	TTTGQA	M11230	579-639	-25.3	Sup(UUA)	TTTRQA	90
Primates	HSA41C122	Z15821	(c) 212-141	-26.0	Ala(UGC)	RSRRNB	88
	HSMMDBC_04	M89651	(c)9806-9723	--[7]	Sec(UCA)	HSTGSS	100
	HSA45B021	Z13399	151-238 (f)	-30.6	Ser(CGA)	LDSERU	76

[1]The average distance between the 3'-end of the B box and the termination site is 20 nt (standard deviation (sd)=13.5, n=231). In cases of limited 3'-sequence information without any detectable termination site, candidate tRNA genes are evaluated on the basis of the intermediate score only (see Experimental Procedures) if less than 47 nt (mean + 2×sd) beyond the B box are available (PTGAL1O, EHRPTARQ, TTTGQA), whereas they are discarded if no transcription termination signal is present within a B box flanking region longer than 47 nucleotides.

[2]In all cases, except EHRPTARQ, predicted anticodons are identical to those of the most similar known tRNA genes. For SCTY109), SCTY4(1), SCTY4(2), GMSB1TUB, HSA41C122 and HSMMDBC_04, the anticodon was predicted after examination of potential secondary structures, whereas in all the remaining cases it was deduced according to the rules described in Experimental Procedures.

[3](f) indicates newly predicted tRNA genes that are also recognized by the algorithm of Fichant amd Burks (11).

[4](c) indicates a tRNA gene found in the complementary strand.

[5](t) indicates *S.cerevisiae* tRNA genes flanked by transposon related sequences.

[6]The numbering (1) and (2), in the case of SCTY4, refers to two distinct potential tRNA genes identified in the same sequence entry.

[7]No total score is assigned to selenocysteine tRNA genes by the standard search program (see Results).

new potential tRNA genes were detected among plant sequences, and the final false-positive rate for this set of sequences corresponds to 0.0083% (Tables 3 and 4).

The scanning of invertebrate sequences led to three false negative predictions out of a total of 285 known tRNA genes (false negatives percentage of 1.04, Table 3). Among the three genes that were not detected, two correspond to a $tRNA^{Leu}$ gene from *D.discoideum* (DDLEU1UAA, DDLEU2UAA) with an unusual A box, 'GTAGGAAAGTCTGGTTAAATCC' (15), in which positions 17, 17a, 20a and 20b are all occupied (19). A similar situation was found in the case of the third gene that was missed (TBTRNA3, $tRNA^{Arg}$ from *Trypanosoma brucei*; 34), which also contains a non canonical A box. Nine new potential tRNA genes were identified in this set of sequences, with a false positive rate of 0.0096% (Tables 3 and 4).

False negative rates were 0 and 1.30% in the case of rodent and other vertebrate sequences, respectively. The one vertebrate tRNA gene that was missed (BTTRSERC, 35), out of a total of 77 known tRNA sequences, lacks a stretch of thymine residues within 133 nucleotides from the 3'-end of the B box. A rodent tRNA gene (RNTRNLEU, 36), out of a total of 83 known sequences, was also initially missed by the algorithm. However, subsequent analysis of the reverse sequence, showed that this apparent failure was due to the fact that this sequence was entered into the EMBL database in a non-standard orientation. No new potential tRNA gene was found in either group, and the corresponding false positive rates were 0.0355% for rodents and 0.0144% for other vertebrates (Table 3).

No false negative instance, out of a total of 78 known tRNA genes, was found in the scanning of primate sequences. Three

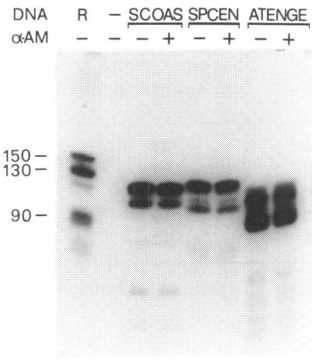

Figure 3. *In vitro* transcription analysis of two potential tRNA genes in SPCEN114 and ATENGE. Transcription reactions were carried out in the presence of a *S.cerevisiae* nuclear extract, and allowed to proceed for only 10 minutes at 20°C in order to minimize primary transcript processing (see Experimental Procedures). Reactions in lanes labeled *SPCEN*, *ATENGE* and *SCOAS* were programmed with PCR-amplified DNA fragments (100 fmol) containing the putative $tRNA^{Leu(CAA)}$ (SPCEN114) and $tRNA^{Leu(UAA)}$ (ATENGE) genes, and the previously predicted (11) $tRNA^{Leu(UAG)}$ (SCOASOAS), respectively. No template was added to the reaction shown in lane labeled (–). Where indicated (+), reaction mixtures were supplemented with α-amanitin, at a concentration (20 μg/ml) known to completely inhibit yeast RNA polymerase II. Transcripts synthesized in three separate reactions programmed with reference tRNA genes (see text for details) were mixed before electrophoresis and used as references (*R*). The migration position and the length of primary transcripts corresponding to $tRNA^{Leu3}$+20 (150 nt), $tRNA^{Leu3}$ (130 nt) and $tRNA^{Arg}$ (90 nt) are shown on the right.

new potential tRNA genes, including one coding for a $tRNA^{Sec}$ (HSMMDBC_04), were identified in this taxonomic group with a rate of false positive predictions of 0.0031% (Tables 3 and 4).

In vitro transcription analysis of new potential tRNA genes

Sequences flanking tRNA genes have previously been shown to influence transcription efficiency (13,37). Upstream control elements are involved, for example, in determining the different transcription rates of distinct members of the same tRNA gene family and in regulating the tissue specific expression of a $tRNA^{Ala}$ gene in *Bombyx mori* (37 and references therein). Although these control elements are functionally important for transcription, their sequences are in most cases very poorly conserved and cannot thus be taken into account by any general search algorithm. A possible way to overcome this limitation and to distinguish between authentic tRNA genes and tRNA derived, but transcriptionally inactive, elements could be to directly test the template activity of newly identified, potential tRNA genes by *in vitro* transcription. We applied this type of experimental analysis to two new potential tRNA genes, SPCEN114 and ATENGE, both comprised in the lower range of similarity, 70% and 74% respectively, with known tRNA genes (Table 4). A third sequence, SCOASOAS, previously identified by computer analysis (11) as a *S.cerevisiae* $tRNA^{Leu(UAG)}$ gene (100% homology with the corresponding tRNA) was used as an internal reference. *In vitro* transcription reactions were carried out in the presence of a *S.cerevisiae* nuclear extract, and were programmed with PCR-amplified DNA fragments comprising the coding region of either gene plus about 100 bp of 5′- and 3′-flanking sequences (see Experimental Procedures). As shown in Figure 3, the transcriptional output of reactions carried out in the presence of a linear DNA fragment corresponding to the $tRNA^{Leu(UAG)}$ gene (lanes labeled *SCOAS*) was comparable to that of parallel reactions supplemented with reference tRNA genes (*R*) contained in negatively supercoiled plasmids. More importantly, *in vitro* transcription reactions programmed with DNA fragments corresponding to the two newly predicted tRNA genes (SPCEN114 and ATENGE) both resulted in the specific, a-amanitin insensitive synthesis of RNA products corresponding in size to either primary or partially processed tRNA transcripts (Figure 3, lanes *SPCEN* and *ATENGE*). We conclude from these data that the $tRNA^{Leu(UAG)}$ gene in SCOASOAS (25), as well as the potential $tRNA^{Leu(CAA)}$ and $tRNA^{Leu(UAA)}$ genes found, respectively, in SPCEN114 (24) and ATENGE (23), are all efficiently utilized as templates by the RNA polymerase III transcription apparatus.

Search of extragenic B boxes in the 3′-flanking region of tRNA genes

Studies in yeast and in insect systems have shown that 3′-flanking sequences of tRNA genes can also contribute to transcriptional modulation (13, 37–39). As in the case of 5′-flanking elements, the sequence identity of these downstream elements is rather ill defined. In the case of a $tRNA^{Ala}$ gene from *B.mori*, however, sequences resembling an imperfect B box are present in a region of the 3′-flanking sequence that contributes both to transcription efficiency and transcription factor binding (39). This finding is not too surprising, since canonical B box sequences are known to be present in the 3′-flanking region of the yeast U6 gene (40), which is also transcribed by RNA polymerase III, and in a large fraction (more than 80%) of tRNA genes from the slime mold *D.discoideum* (15). The functional significance of these extragenically located B boxes is not yet fully understood, particularly in the case of *D.discoideum*, for which an homologous *in vitro* transcription system is not available. We were thus interested to know whether this motif is a unique feature of some particular organisms only or whether it is present in tRNA genes from a larger variety of organisms. To address this point, well characterized extragenic B boxes of *D.discoideum* tRNA genes were used as a reference system (15). An extragenic B box searching subroutine was developed and used, in combination with the standard program, for the analysis of *D.discoideum* tRNA genes. All known extragenic B boxes, including some that are tandemly arranged, were correctly recognized at distances from the intragenic B box ranging from 37 to 50 nucleotides in the case of single elements and up to 94 nucleotides in the case of double repeated elements. We then searched for the presence of extragenic B boxes, at a maximum distance of 100 nucleotides from the 3′-end of the corresponding intragenic element, in all known and new potential tRNA genes. At a cut-off value of −9.30, which includes no more than 92.6% of all the B boxes in our sample set (Figure 2, inset), five new extragenic B boxes were identified in 4 different taxonomic groups (Table 5); two of them belong to newly predicted tRNA genes (ATENGE, SCPP1A), whereas the remainder were found associated to known tRNA genes. Two to three complementary base pairs, with the potential ability to form a stem region, are present on both sides of these extragenic elements, which share varying degrees of similarity (from complete to 7 out of 11 nucleotides) with their corresponding intragenic B boxes. Interestingly, an identical extragenic B box sequence is present, at nearly the same distance from the intragenic B box, in two allelic forms (SCPP1A, SCC90A; 41) of the same yeast $tRNA^{His}$ gene.

Table 5. Extragenic B boxes in the 3'-flanking region of tRNA genes

Taxonomic group	Entry name	Accession number	Positions[1]	Nucleotide sequence	B box score	Distance from the intragenic B box (bp)[2]
Fungi	SCC90A	X17306	2275 (c)	GGTTCAAAGCC	-8.2	89
	SCPP1A	M77175	2836 (c)	GGTTCAAAGCC	-8.2	87
Plants	ATENGE	X58107	4543	GGTTCGAACCC	-3.0	92
Invertebrates	DMRNA2	V00237	1946	GGTTCAAAGCC	-8.2	46
Vertebrates	XLTRNA	M32259	1504	GGTTTAAATCC	-8.6	71

[1]Reported positions correspond to the first nucleotide of the extragenic B box; (c) indicates an extragenic B box associated to a tRNA gene in the complementary strand.
[2]The distance from the 3'-end of the intragenic B box is shown.

DISCUSSION

A multistep weight matrix analysis of the nucleotide sequence and of the relative positioning of transcriptional control elements forms the basis of the present tRNA gene search algorithm. By relying exclusively on linear sequence analysis, our algorithm has the ability to recognize even structurally unusual tRNAs and it can be adapted to the search of special classes of tRNA genes (e.g tRNA^{Sec} genes), as well as to the detection of particular sequence motifs (e.g. extragenic B boxes). On the other hand, a potential shortcoming of this linear sequence approach, as compared for example to search methods mainly based on secondary structure analysis, is the difficulty to directly distinguish, at least in some cases, between true tRNA genes and transcriptionally active tRNA-derived elements. The different performance of these two types of approach, as applied to the scanning of eukaryotic nuclear DNA sequences, is best illustrated by a direct comparison between our results and those generated by the exhaustive search algorithm previously developed by Fichant and Burks (11). The rate of false negative predictions yielded by the latter algorithm, 2.5% on a total of 432 known eukaryotic tRNA genes, is about 3 times higher than the corresponding rate we obtained from the analysis of 61580 nuclear DNA sequences containing a total of 940 known tRNA genes. The opposite situation was found in the case of false positive predictions, whose rate, for the algorithm of Fichant and Burks (11), was about 4-fold lower than ours (0.0033% versus 0.0137%). More noteworthy, however, is the fact that 6 out of 7 genes that are not recognized by our algorithm are correctly identified by the algorithm of Fichant and Burks (11) and that, conversely, known tRNA genes that escaped detection by the latter algorithm, all display total score values above the threshold when analyzed with our search method. In fact, the combined use of the two algorithms leads to a total false negative rate that is very close to zero (1 missed tRNA gene on an updated test set of 940 known tRNA sequences), and their nearly perfect complementarity most likely results from the utilization of different diagnostic features. Interestingly, the only gene that is missed by both algorithms codes for a promiscuous *T.brucei* tRNA (TBTRNA3) that is present both in cytosol and mitochondria (34). Rates of false negative predictions yielded by our algorithm did not vary by more than 2-fold between different taxonomic groups (Table 3), an indication of the general applicability of our method to the search of nuclear eukaryotic tRNA genes. In contrast, the rate of false positives in different sets of sequences was found to be highly variable: it ranged from 0.0006% in fungi to 0.0355% in rodents. The rate of false positive predictions in different taxonomic groups does not vary

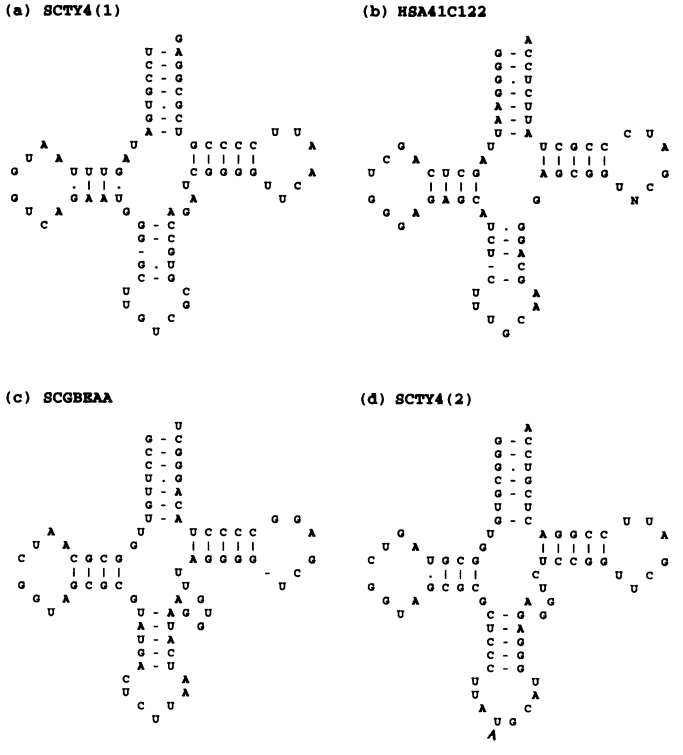

Figure 4. Potential secondary structures of newly predicted tRNAs found in: (a) SCTY4(1), (b) HSA41C122, (c) SCGBEAA, (d) SCTY4(2). (−) indicates an empty position, (^) indicates an extra nucleotide; (see Text and Table 4 for details).

in a random fashion, however, and it is somehow related to the presence and relative abundance of interspersed elements originated from tRNA genes (42). It is highest in the case of genomes, like for instance those of rodents, that are known to contain tRNA-derived repetitive elements (e.g. rat ID, mouse B2; 42), and lowest in others, like fungi, where the presence of similar elements has never been detected. In keeping with this observation, the ratio between the number of false positives yielded by the scanning of a given set of sequences and those found in a random permutation of the same set of sequences varies from 37 in rodents to 1 in fungi (Table 3). More specifically, 149 false positive instances—76 rat ID sequences, 49 mouse B2 and 1 B1 sequences, 9 Alu type III-like sequences, 1 unrelated tRNA-like element and 9 pseudogenes—were found in the scanning of rodent DNA sequences, whereas only 4 false positives

were obtained from a random permutation of this same set of sequences. On the other side, only one false positive was obtained from either type of analysis in fungi. Intermediate false positive rates were obtained for the other taxonomic groups, where a moderate excess of false positives in real sequences with respect to randomized sequences was due to either pseudogenes and/or known tRNA-like repetitive elements (Table 3). Noteworthy, a first case of a plant tRNA-derived repetitive element has recently been discovered in the tobacco genome (43). The 10 false positives we found in plant sequences (half of them in *Arabidopsis* and sunflower) thus appear to be particularly interesting, and their further characterization might expand the repertoire of tRNA-related repetitive elements in plants.

The scanning of the eukaryotic nuclear DNA database with our search algorithm yielded thirty new sequences with total score values above the threshold and exhibiting significant similarity with known tRNA genes. Other criteria, besides coding sequence similarity, support the idea that these sequences may correspond to authentic tRNA genes. These include, for example, the ability of two of them (ATENGE and SPCEN114; 23,24) to efficiently support *in vitro* transcription reactions (Figure 3), and the characteristic presence in close proximity to a number of new potential tRNA genes from *S.cerevisiae* of transposable elements (44), (Table 4). Additional support was provided by a comparison of the nucleotide composition of upstream sequences (up to −50) of our predicted tRNA genes with the corresponding region of 576 known tRNA genes. In fact, important interactions between tRNA genes and components of the RNA polymerase III transcription machinery are known to take place in this region, which has a characteristic A/T rich composition (37,45). As a result of this test, we found that the average base composition (and the relative standard deviation) for our candidate tRNA genes (A=37%, 10.1; T=32%, 8.0; G=15%, 6.4; C=16%, 6.8) is remarkably close to that of the 576 sample tRNA genes (A=33%, 10.0; T=31%, 8.4; G=18%, 9.1; C=18%, 7.3).

New potential tRNA genes recognized by our algorithm, with the exclusion of two incomplete sequences (PTGAL10, TTTGQA) and a potential $tRNA^{Sec}$ gene (HSMMDBC_04), were then cross-analyzed using the algorithm of Fichant and Burks (11). This analysis confirmed our predictions for 18 out of 27 candidate sequences (see Table 4). While further strenghtening the notion that these 18 sequences do indeed correspond to authentic tRNA genes, this finding also raises a question as to the reason why the remaining 9 sequences were recognized by our algorithm only. A more detailed analysis of these nine sequences showed that seven of them (SCGBEAA, SCTY4(1), SCTY4(2), SPCEN114, GMSB1TUB, CEUNC33G, HSA41C122) exhibit localized deviations from the standard cloverleaf model of secondary structure (1 or 2 positions) which account for the apparent discrepancy between the two algorithms. For example, single nucleotides appear to be missing from either the anticodon arm of SCTY4(1) and HSA41C122 or from the TΨC loop of SCGBEAA, whereas an extra nucleotide is present in the anticodon loop of SCTY4(2). If one takes into account these localized imperfections, the resulting structures of all these elements closely conform to a standard cloverleaf model of secondary structure (Figure 4). Significant deviations from the standard model of secondary structure have previously been found also in the case of known nuclear tRNA genes; for instance, the secondary structure of a yeast frameshift suppressor tRNA gene, with an extranucleotide in the anticodon loop (46), closely resembles the predicted structure of SCTY4(2). Sequencing imprecisions might also explain at least some of these seven particular cases, which we tend to consider, nevertheless, as representative of true tRNA genes. A different situation was found in the remaining two cases (SCTY109, EHDNA), where striking structural anomalies, mainly within the acceptor arm, are present; both sequences probably correspond to tRNA pseudogenes, and were likely generated from *S.cerevisiae* $tRNA^{Trp}$ (SCTRWSIG) and *E.hystolitica* $tRNA^{Glu}$ (EHTRGLUG) genes, respectively, through a duplication event (Table 4).

Some of the newly identified tRNA genes were found in rather uncommon positions within the nuclear genome. Two of them are situated in centromeric regions: SCCEN11D01 [$tRNA^{Leu(CAA)}$] represents the first case of a centromeric tRNA gene in *S.cerevisiae*, whereas SPCEN114 is an additional member of a cluster of tRNA genes that is present in the centromeric region of chromosome II of fission yeast (3,4). Very peculiar is also the localization of three new tRNA genes [CEGPA1/$tRNA^{Leu(CAG)}$, CEPGPAG/$tRNA^{Lys(CTT)}$, CEUNC33G/$tRNA^{Arg(ACG)}$] that were found within introns. Interestingly, all of them belong to *Caenorhabditis elegans*, an organism in which three additional intronic tRNA genes have recently been revealed by the sequencing of 2.2 megabases of genomic DNA (R.Wilson, personal communication). This finding is reminiscent of the sequences coding for small nucleolar RNAs that have been identified in the introns of various ribosomal protein genes (47,48, and references therein). Most of the other tRNA genes were found at various distances, either 5′ or 3′, from protein coding genes. Minimum distances were 248 bp upstream of the TATA box in the case of ATCPYLP (*A.thaliana* $tRNA^{Tyr}$/carboxypeptidase Y-like protein genes) and 30 bp downstream of the polyadenylation signal in the case of SCGBEAA (*S.cerevisiae* glycogen branching enzyme/$tRNA^{Lys}$ genes).

The identification of extragenic B boxes in tRNA genes from various organisms, other than *D.discoideum* (15), indicates that this motif is likely to be more general than previously thought. Sequences downstream of the transcription termination site have been shown to influence tRNA gene transcription in some systems (37−39), apparently by participating in stable complex formation with transcription factors, and TFIIIC binding to the extragenic B box of the yeast U6 gene has recently been found to relieve chromatin repression of transcription (49). The exact functional significance of this sequence motif is not yet fully understood, however. By providing new experimentally approachable cases, for example the two yeast $tRNA^{His}$ genes (Table 5), our findings will allow a more extensive investigation of the functional role of these extragenic elements.

Key features of the present searching approach—namely, its ability to examine a composite genetic signal made of multiple, heterogeneously spaced regions—should allow to extend its application to the analysis of other multipartite regulatory elements. (An updated compilation of aligned A and B box sequences from 940 eukaryotic nuclear tRNA genes, and the computer search program are freely available on request.)

Entry names and accession numbers of various tRNA sequences mentioned in the text: ATPAT1 (X14103), ATPATY3 (X54370), BTTRSERC (X51356), CEARGTRNA (X51770), CETGKA (K01852), DDLEU1UAA (X59572), DDLEU2UAA (X59573), DMRP (K02463), DMTGL (M15822), DMTGYC (M21613), EHTRGLUG (Z11506), HSTGSS (K02923), HSTRNS1 (X06760), LDSERU (X13888), MCTRFM (X16759), NRTY8 (X58508), PCTRNA (X68436),

RNTRNLEU (X00710), RSRRNB (X53854), SC1 (Z11113), SCANBI (J05455), SCBAR1A (J03573), SCCYP1 (X13793), SCD (K00171), SCFT5 (J01369), SCGUT2P (M74328), SCHAP1 (J03152), SCHYP2 (X56236), SCIP (K02963), SCK1 (K00286), SCLEURNA (X56506), SCLSC (M26843), SCMCM3G (X53540), SCOASOAS (X04493), SCRN15 (V01330), SCSM (K00368), SCSTE6G (X15428), SCTGHC (K01597), SCTGL (K01599), SCTIF51A (M63541), SCTRNCYS (X01939), SCTRWSIG (M15249), SCTVTY3 (X55335), SPTGSUP (K01631), TATRTY1 (X51731), TBTRNA3 (X57046), TTTRQA (M11464), TVTRNS (X60931), YSTRYP (M10721).

ACKNOWLEDGEMENTS

We are grateful to Professor Gian Luigi Rossi for support and encouragement. We thank Drs Gwennaele Fichant and Christian Burks for providing us a copy of their tRNA gene search program, Dr Mitsuhiro Yanagida for plasmid pSS170 and for helpful suggestions concerning the secondary structure of SPCEN114, Dr Dominique Van Der Straeten for the *Arabidopsis* enolase clone, Dr Yolande Surdin-Kerjan for plasmid pEMBLYr25-MET25, Dr Benjamin Hall for plasmid pSArg, Dr Jerry Johnson for plasmids pJD137 and pJD110 and Dr Richard Wilson for communicating results prior to publication. We also acknowledge the Interuniversitary Consortium for Biotechnologies (C.I.B.) for oligonucleotide synthesis and PCR facilities. This work was supported by the National Research Council of Italy (Target Project on Biotechnology and Bioinstrumentation) and by the Ministry of University and Scientific and Technological Research.

REFERENCES

1. Staden, R. (1990) in *Methods in Enzimology* (Doolittle, R.F.,Ed.) **Vol 183**, pp. 193–211, Academic Press
2. Stormo, G.D., *ibidem*, pp. 211–221
3. Khun, R.M., Clarke,L. and Carbon, J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1306–1310
4. Takahashi, K., Murakami, S., Chikashige, Y., Niwa, O. and Yanagida, M. (1991) *J. Mol. Biol.* **218**, 13–17
5. Andreadis, A., Hsu, Y.P., Kohlhaw, G.B. and Schimmel, P. (1982) *Cell* **31**, 319–325
6. Suter, B. and Kubli, E. (1988) *Mol. Cell. Biol.* **8**, 3322–3331
7. Papanicolau, C., Gouy, M. and Ninio J. (1984) *Nucleic Acids Res.* **12**, 31–44
8. Staden, R. (1980) *Nucleic Acids Res.* **8**, 817–825
9. Shortridge R.D., Pirtle I.L. and Pirtle R.M. (1986) *Comp. Applic. Biosci.* **2**, 13–17
10. Marvel, C.C. (1986) *Nucleic Acids Res.* **14**, 431–435.
11. Fichant G.A. and Burks C. (1991) *J. Mol. Biol.* **220**, 659–671
12. Sharp, S., Schaack, J., Cooley, L., Burke, D. and Soll, D. (1985) *Crit. Rev. Biochem.* **19**, 107–144
13. Geiduschek, E.P. and Tocchini-Valentini, G.P. (1988) *Annu. Rev. Biochem.* **57**, 873–914
14. Staden R. (1984) *Nucleic Acids Res.* **12**, 507–521
15. Hofmann, J., Schumann, G., Borschet, G., Gosseringer, R, Bach, M., Bertling, W.M., Marschalek and Dingermann, T. (1991) *J. Mol. Biol.* **222**, 537–552
16. Higgins D.G., Fuchs R., Stoehr, P.J. and Cameron G.N. (1992) *Nucleic Acids Res.* **20**, Suppl:2071–2074
17. Burks, C., Cinkosky M.J., Fischer W.M., Gilna P., Hayden J.E.D., Keen G.M., Kelly M., Kristofferson D. and Lawrence J. (1992) *Nucleic Acids Res.* **20**, Suppl: 2065–2070
18. Krupp J.L, Shu, H.H. and Martin, N.C. (1988) *Nucleic Acids Res.* **16**, 770
19. Schimmel R, Soll D. and Abelson J.N. (1979) in *Transfer RNA: Structure, Properties and Recognition*, Cold Spring Harbour Laboratory, N.Y., 518–519
20. Lipman, D.J. and Pearson, W.R. (1985) *Science* **227**, 1435–1441
21. Myiers E.W. and Miller W (1988) *Comp. Applic. Biosci.* **4**, 11–17
22. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) *Science* **239**, 487–491
23. Van der Straeten D., Rodriguez-Pousada R.A., Goodman H.M. and Montagu M. (1991) *Plant Cell* **2**, 719–735
24. Chikashige Y, Kinoshita N., Nakaseko Y., Niwa O. and Yanagida, M. (1989) *Cell* **57**, 739–751
25. Kerjan P., Cherest H. and Surdin-Kerjan Y. (1986) *Nucleic Acids Res.* **14**, 7861–7871.
26. Stange, N., Beier, D. and Beier, H. (1992) *Eur. J. Biochem.* **210**, 193–205
27. Dieci G., Duimio L., Coda-Zabetta, F., Sprague, K.U. and Ottonello, S. (1993) *J. Biol. Chem.* **268**, 11199–11207
28. Raymond, G.J. and Johnson, J.D. (1983) *Nucleic Acids Res.* **11**, 5969–5988
29. Baker, R.E. and Hall, B.D. (1984) *EMBO J.* **3**, 2793–2800
30. Straby, K.B. (1988) *Nucleic Acids Res.* **16**, 2841–2857
31. Munz, P., Amstutz, H., Kohli, J. and Leupold, U. (1982) *Nature* **300**, 225–231
32. Newmann, A.J., Ogden, R.C. and Abelson, J. (1983) *Cell* **35**, 117–125
33. Szweykowska-Kulinska, Z., Jarmolowski, A. and Augustyniak, J. (1989) *Gene* **77**, 163–167
34. Mottram, J.C., Bell S.D., Nelson R.G. and Barry D. (1991) *J. Biol. Chem.* **266**, 18313–18317
35. Chee, M.S., Rizos, H., Henderson, B.R., Baker, R. and Stewart, T.S. (1991) *Mol. Gen. Genet.* **231**, 106–112
36. Rosen, A., Sarid, S., Daniel, V. (1984) *Nucleic Acids Res.* **12**, 4893–4906
37. Sprague, K.U. (1993) in *Transfer RNA* (Soll, D. and RajBhandary, U., eds.), ASM, in press
38. Allison, D.S. and Hall, B.D. (1985) *EMBO J.* **4**, 2657–2664
39. Young, L.S., Rivier, D.H. and Sprague, K.U. (1991) *Mol. Cell. Biol.* **11**, 1382–1392
40. Brow, D.A. and Guthrie, C. (1990) *Genes Dev.* **4**, 1345–1356
41. Lochmuller, H., Stucka, R. and Feldmann, H. (1989) *Curr. Genet.* **16**, 247–252
42. Okada, N. (1991) *Current Opinion in Genetics and Development* **1**, 498–504.
43. Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. and Machida, Y. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 6562–6566
44. Sandmeyer, S.B., Hansen, L.J. and Chalker, D.L. (1990) *Annu. Rev. Genet.* **24**, 491–518
45. Leveillard, T., Kassavetis, G.A. and Geiduschek, E.P. (1993) *J. Biol. Chem.* **268**, 3594
46. Cummins, C.M., Donahue, T.F. and Culbertson, M.R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3565–3569
47. Kiss, T. and Filipowicz (1993) *EMBO J.* **12**, 2913–2920
48. Fragapane, P., Prislei, S., Michienzi, A., Caffarelli, E. and Bozzoni, I. *EMBO J.* **12**, 2921–2928
49. Burnol, A.-F., Margottin, F., Huet, J., Almouzni, G., Prioleau, M.-N., Méchali, M. and Sentenac, A. (1993) *Nature* **362**, 475–477