J Mol Evol (2007) 64:248–260
DOI: 10.1007/s00239-006-0135-6

JOURNAL OF MOLECULAR EVOLUTION

Looking for Organization Patterns of Highly Expressed Genes: Purine-Pyrimidine Composition of Precursor mRNAs

A. Paz, D. Mester, E. Nevo, A. Korol

Institute of Evolution, Haifa University, Mount Carmel, Haifa 31905, Israel

Received: 21 June 2006 / Accepted: 19 November 2006 [*Reviewing Editor:* Dr. Axel Meyer]

Abstract. We analyzed precursor messenger RNAs (pre-mRNAs) of 12 eukaryotic species. In each species, three groups of highly expressed genes, ribosomal proteins, heat shock proteins, and amino-acyl tRNA synthetases, were compared with a control group (randomly selected genes). The purine-pyrimidine (R-Y) composition of pre-mRNAs of the three targeted gene groups proved to differ significantly from the control. The exons of the three groups tested have higher purine contents and R-tract abundance and lower abundance of Y-tracts compared to the control (R-tract—tract of sequential purines with $R_n \geq 5$; Y-tract—tract of sequential pyrimidines with $Y_n \geq 5$). In species widely employing "intron definition" in the splicing process, the Y content of introns of the three targeted groups appeared to be higher compared to the control group. Furthermore, in all examined species, the introns of the targeted genes have a lower abundance of R-tracts compared to the control. We hypothesized that the R-Y composition of the targeted gene groups contributes to high rate and efficiency of both splicing and translation, in addition to the mRNA coding role. This is presumably achieved by (1) reducing the possibility of the formation of secondary structures in the mRNA, (2) using the R-tracts and R-biased sequences as exonic splicing enhancers, (3) lowering the amount of targets for pyrimidine tract binding protein in the exons, and (4) reducing the amount of target sequences for binding of serine/arginine-rich (SR) proteins in the introns, thereby allowing SR proteins to bind to proper (exonic) targets.

Key words: Purine tracts — Pyrimidine tracts — Splicing efficiency — Highly expressed genes — Intron definition — Exon definition

Introduction

Messenger RNAs (mRNAs) of viruses and species from all the three domains of life generally tend to be purine (R)-rich (Szybalski et al. 1966; Smithies et al. 1981; Bell and Forsdyke 1999; Lao and Forsdyke 2000; Paz et al. 2004). The reason for this phenomenon is not well understood. The "politeness" hypothesis (Zuckerkandl 1986) assumes that R-loading in mRNAs may hinder the formation of "forbidden" double-stranded RNA (dsRNA), which severely disrupt translation and trigger various intracellular alarms (Forsdyke 1999). During initiation of translation, i.e., the rate-limiting step to protein synthesis (van der Velden et al. 2002; Arava et al. 2003), hairpin structure in the mRNA causes the 40S ribosomal subunit to stall, and thus translation is slowed down or even stopped (van der Velden et al. 2002; Kozak 2005). From first principles, the formation of dsRNA by (inter- and intrastrand) Watson–Crick base-pairing might be lower in R-loaded mRNAs than in mRNAs with equal amounts of purines and pyrimidines. Thermophilic prokaryotes have a high purine content and R-tract abundance in their mRNA, compared to mesophilic species (R-tract—tract of sequential purines with $R_n \geq 5$) (Paz et al. 2004). It was hypothesized that the high R-bias of

Correspondence to: A. Korol; *email:* korol@research.haifa.ac.il

the thermophiles mRNA is a thermoadaptation (Lao and Forsdyke 2000; Lobry and Chessel 2003; Paz et al. 2004). Short-range purine-pyrimidine sequence composition was found to be an important characteristic of species genome organization, on the above-gene level (Paz et al. 2005).

In both thermophilic and mesophilic prokaryotes, mRNAs of ribosomal proteins (RPs) and heat shock proteins (HSPs) have significantly higher R-bias and R-tracts abundance than their species average (Paz et al. 2004). Similarly, mRNAs of prokaryotic aminoacyl tRNA synthetases (ARSs) also have higher R-bias and R-tract abundance than their species average (A. Paz, unpublished results). We want to check whether in the eukarya, as is the case in the prokarya, the mRNAs of three groups of highly expressed genes, RPs, HSPs, and ARSs, have higher R-bias and R-tract abundance than the species average. The major difference between the maturation of mRNA in eukarya compared to prokarya, i.e., the splicing of precursor mRNA in the eukarya, prevents direct extrapolations and expectations for similarity. This is because there might be a need for splicing signals in the exons (besides the *cis*-component sequences located in the introns needed for splicing), which can influence the R-Y composition of the exons. The central importance of the splicing process to the expression level of genes in eukarya poses additional interesting questions: Does the R-Y composition of introns of the three targeted groups differ from the species average? If indeed the R-Y composition of both exons and introns of the three targeted groups differ from the average, then how can these differences be attributed to the higher expression of the former gene groups compared to the expression of "average" genes? (One can speculate that a specific R-Y composition of pre-mRNA might enable a high rate and efficiency of splicing.)

In the current study we analyze precursor RNA (pre-mRNA) of 12 eukaryotic species, regarding R-Y (content and distribution) within exons and introns. We compared the three protein-coding gene groups, RPs, HSPs, and ARSs, with a control group of randomly selected protein-coding genes. Based on the literature (see below), the expression level of the three targeted groups of genes is higher than the expression of "average" genes of the species. Thus, we referred to the former genes as targeted highly expressed genes (THEGs). An additional feature that these genes share is their cooperation in protein synthesis and maturation.

High expression of genes that belong to the three targeted groups was reported in expression sequence tag (EST) studies of various eukaryotes, including the unicellular yeast *Saccharomyces cerevisiae* and multicellular species (Herruer et al. 1987; Warner 1999; Seshaiah and Andrew 1999; Warrington et al. 2000; Hsiao et al. 2001; Yu et al. 2001; van Ruissen et al. 2002). Fifty percent of the estimated RNA-polymerase II-mediated transcription initiation events in the yeast involve RP genes (Warner 1999). In many tissue types of the metazoan, the expression of RPs is high (ranked within the highest third), possibly to fit the requirements of a high amount of ribosomes during elevated protein synthesis. HSPs and ARSs are also involved in the process of protein synthesis and maturation. The expression levels of many HSP genes are not exclusively related to stress, e.g., high temperature. Indeed, many HSP chaperones are expressed constitutively. In *Saccharomyces cerevisiae* and mammals, various HSP chaperones, possibly including the nascent polypeptide-associated complex, interact with ribosomes in processing and in protecting nascent polypeptides exiting the ribosome (Fewell et al. 2001; Frydman 2001; Hartl and Hayer-Hartl 2002). Eukaryotic ARSs are involved in noncanonical (noncatalytic) but very important functions related to proteins synthesis (Lee et al. 2004; Park et al. 2005). In particular, some ARSs form a complex with the translation elongation factor EF-1H complex (Bec et al. 1994; Sang Lee et al. 2002). This complex has also been shown to functionally interact with EF-1a (Negrutskii et al. 1999). The proposed role of this complex is to facilitate the delivery of the charged tRNA to the ribosome.

The Main Questions of the Study

The three main questions of the study are as follows.

1. Is there a trend of eukaryotic mRNAs of RPs, HSPs, and ARSs to have higher than average R-bias and R-tracts abundance?
2. Do the introns of these gene groups differ from average genes in the R-Y composition? If the answers to the two questions are positive, then how are the differences in the R-Y composition attributed to the higher level of expression of the former gene groups compared to "average" genes?
3. Are there any differences between lower and higher eukaryotes in the patterns of sequence organization, with respect to the R and Y content and homotract distribution in pre-mRNAs of THEG genes, "average" genes, or both?

We expected that THEG mRNA will have higher than average R-bias and higher than average abundance of R-tracts. These features (if they exist) might contribute to both a higher efficiency of splicing and a reduction of translation disturbances. These two roles of R-biased mRNA do not exclude the coding role of the purine-rich sequences, contributing to an increased level of charged amino acids, needed for THEG protein functions. We also had preliminary assumptions about the R-Y composition of the introns: according to Chargaff's second parity rule (Karkas et al. 1968; Rudner et al. 1968), there is an approximate equality in the nucleotide content of single-stranded DNA (%A = %T and %C = %G). This means that there should also be equality of %R to %Y (as R = A + G and Y = T + C). Although this rule was confirmed in a long-range analysis of single-stranded DNA of many species (Prabhu 1993), the aforementioned purine bias of mRNA (Szybalski et al. 1966; Smithies et al. 1981; Bell and Forsdyke 1999; Lao and Forsdyke 2000; Paz et al. 2004) means that in short range, there are deviations from this parity rule. It is interesting to check whether within pre-mRNA sequences there is a trend for "compensation" of large deviations from Chargaff's second parity rule within the exons via specific nucleotide content of introns. This question might be especially relevant to the introns of THEGs, if our assumption that their mRNAs are highly R-biased is indeed correct. Therefore, introns are pyrimidine biased in general, and we hypothesize that this bias may be higher in genes with exceedingly purine-biased mRNA and that the contrasted R-Y-biases of THEG exons and introns contribute to high expression of THEGs.

Materials and Methods

We analyzed pre-mRNAs of 12 eukaryotic species: *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Anopheles gambiae, Apis mellifera, Ciona intestinalis, Danio rerio, Takifugu rubripes, Xenopus tropicalis, Gallus gallus, Mus musculus*, and *Homo sapiens*. For the 11 metazoan species, the sequence and annotation data were retrieved from the Ensembl Genome Browser (http://www.ensembl.org/index.html). For the plant *Arabidopsis thaliana* we used gene models and annotations of TAIR (http://www.arabidopsis.org/index.jsp) and MIPS (http://www.mips.gsf.de/proj/thal).

Gene selection. The analyzed genes belong to two groups: THEGs (1041 genes) and genes used as control (1200 genes). The genes of the control group were selected randomly from the annotated databases of the 12 species. We use the abbreviation RSGs (randomly selected genes) for this group. Only intron-containing genes were selected. In each of the 12 species, the total number of analyzed RSGs was at least equal to the number of THEGs.

THEGs. The THEG group included five subgroups of genes, encoding RPs, mitochondrial RPs (mtRPs), HSPs, HSPs with a prokaryotic-like HSP40 signature (DNAJs), and ARSs.

RPs and mtRPs. The RP group has a larger number of annotated genes than other THEG groups. The number of predicted ribosomal proteins is approximately 80 (although some species might have < 80 RPs); for the mitochondrial RPs the predicted number is smaller. When more than one annotated gene encoding for a certain RP was found in the database for a species, we arbitrarily chose only one of these genes.

HSPs. This group was classified into six protein families on the basis of molecular mass. These included (1) high molecular weight HSPs (100 kDa); (2) the HSP90 family; (3) the highly conserved HSP70 family, which represents the most prominent eukaryotic group of HSPs; (4) the HSP60 family of chloroplasts and mitochondria; (5) eukaryotic homologues of *E. coli* DnaJ (HSP40); and (6) the family of small HSPs, members of which are expressed predominantly in plants (Caplan et al. 1993).

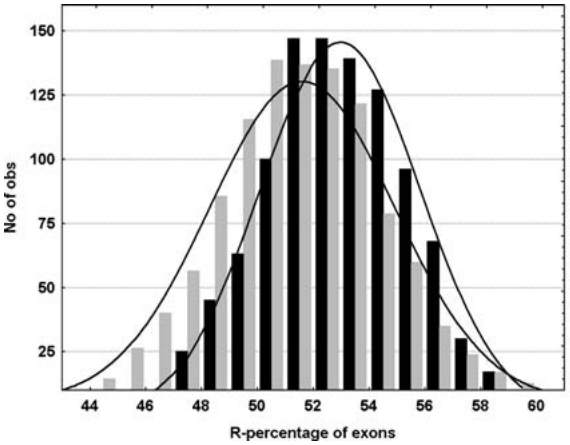

Fig. 1. Distribution of purine content in exons of the targeted groups of highly expressed genes (THEGs) and randomly selected genes (RSGs; control). R-percentage, purine percentage; No. of obs, number of observed exons with the indicated percentage of purines. Black columns: exons of ribosomal proteins, heat shock proteins, and aminoacyl tRNA synthetases (THEG group). Gray columns: exons of RSGs.

ARSs. This is the smallest group; in each species there are only 21 enzymes that belong to this group. All the annotated genes of HSPs and ARSs within the species data were included because the relatively small size of this group compared to the RP group. If several alternative transcripts were presented in the database, we arbitrarily chose only the first suggested annotation found in the database site.

For the analysis of the nucleotide content and sequence organization of pre-mRNA, including purine and pyrimidine tracts, we used our own C^{++} program (ART-5). The statistical calculations were conducted using the STATISTICA 6 program.

Results and Discussion

The main results of our analysis can be concisely summarized as follows. (1) THEG mRNAs have significantly higher purine bias and higher percentages of purines in R-tracts than the control group of genes (Fig. 1). THEG mRNAs also have lower percentages of pyrimidine in Y-tracts than the control (Y-tract—tract of sequential pyrimidines with $Y_n \geq 5$). (2) The level of purine bias in mRNA of eukaryotic species does not seem to be correlated with the optimum growth temperature (OGT) of the species (for poikilotherms) or the body temperature (for homeothermic species). (3) Purine bias of "average" mRNAs in mammals is the lowest among the tested eukaryotes (it is reflected in high proportions of pyrimidines in tracts within exons). (4) THEG introns display a higher pyrimidine bias than the species average in which the splicing of (many or most) pre-mRNAs is based on "intron definition" (the invertebrates and *A. thaliana*). This trend weakens toward the vertebrates and disappears, or even becomes replaced by an opposite trend, in species with a high rate of alternative splicing (mammals and chickens). (5) In all examined species, the introns of THEGs have a lower abundance of R-tracts compared to the control.

THEG Subgroups Have a Higher Exonic Purine Content, Higher Abundance of Pure R-Tracts, and Lower Abundance of Pure Y-Tracts, Compared to Control Groups

We confirmed the already known R-bias of mRNA (Szybalski et al. 1966; Smithies et al. 1981; Bell and Forsdyke 1999; Lao and Forsdyke 2000; Paz et al. 2004). This trend was found in each of the THEG subgroups as well as the RSG groups (Table 1, column 2). In accordance with our expectation (see Introduction), the average purine content in mRNA of THEG genes was found to be higher than that in RSGs (see Table 1, columns 2 and 3, for detailed data and Table 2, columns 3 and 4, for summarized data). As shown in Table 1 (columns 2 and 3), this trend was found in 40 of 43 cases (93%) of comparisons between the THEG subgroups and the RSG average. For the overall trend, the deviation from H_0 of no trend is significant using the Mann-Whitney *U* test at $p < 10^{-6}$, albeit this difference was significant in only 21 of the individual cases (49%). The small number of annotated THEGs of *Ciona intestinalis*, *Takifugu rubripes*, *Xenopus tropicalis*, and *Gallus gallus* seems to be the cause of the nonsignificant difference between these genes and the control group. mRNAs of only two THEG subgroups, the RPs of *D. melanogaster* and *A. thaliana*, display significantly *lower* purine bias than those of the control group (see below). As also shown in Table 1 (columns 4 and 5), in 42 of 43 cases (98%), the mRNAs of THEG subgroups have higher percentages of purines within tracts compared to the control ($p < 10^{-6}$). In individual comparisons, this difference was significant at $p < 0.05$ in 35 of 43 cases. Columns 6 and 7 in Table 1 show that, compared to the control group, the mRNAs of THEG subgroups tend to have lower percentages of pyrimidines within Y-tracts (significance of the overall test, $p < 10^{-6}$); in the individual cases this difference is significant ($p < 0.05$) for only 16 of 43 subgroups. Correspondingly, the ratio (% nucleotides in R-tracts/% nucleotides in Y-tracts) is higher in THEG exons compared to RSG exons (control) in all the species examined, except *D. melanogaster* (Fig. 2).

The mRNAs of *D. melanogaster* and *A. thaliana* RPs display significantly *lower* purine bias than those of the control group. For both species, this is mainly because of the relatively high pyrimidine content of the 5′ untranslated regions (UTRs), whereas the coding regions still have a significantly *higher* purine content than the control group (data not shown). The excess of pyrimidines in Y-tracts within the *D. melanogaster* RPs mRNAs is also significantly higher than that of the control (Table 1, underlined). The 5′ UTRs of *D. melanogaster* RPs mRNAs include a pyrimidine tract with a length of 6–13 bases (not shown). These terminal oligopyrimidine tracts (TOPs) are involved in regulation of translation of RPs (Meyuhas 2000; Marygold et al. 2005). Mammals and other vertebrates also have 5′ TOP within the UTR of the RPs (Meyuhas 2000; Perry 2005). According to our results, within the vertebrates, the R-biases of both the coding and the noncoding regions of the RPs mRNAs override the small local pyrimidine bias caused by the 5′ TOP, resulting in an overall purine bias larger than that of the RSG mRNA.

The relatively low R content of *A. thaliana* RPs mRNAs is mainly displayed as an excess of C over G in the 5′ UTR. These results are in accordance with previous reports on GC-skew (excess of C over G) within the *A. thaliana* genes near the transcription start sites penetrating into the first exons that might be a transcription signal. It is larger in highly expressed genes compared to genes with low expression (Tatarinova et al. 2003; Fujimori et al. 2005).

Species Ecological Conditions and Purine Content in mRNA

As shown in Table 2 (columns 3 and 4), purine content in mRNA is not correlated with the species OGT of the poikilothermic species or the body temperature of the homeothermic vertebrates. However, it is noteworthy that the bee *A. mellifera* displays higher purine content within its mRNA compared to the three other analyzed invertebrates. The difference is significant ($p < 10^{-6}$ using Mann-Whitney *U* test) for both the THEG and the RSG groups. The temperature inside the natural hives of *A. mellifera* is high (32°–36°C [Tautz et al. 2003]) and sometimes can reach 40°C. These temperatures are much higher compared to the OGTs of *C. elegans*, *D. melanogaster*, and *A. gambiae* (20°, 22°, and 26°C, respectively). We speculate that the difference between the bee and the other three invertebrates in the preferable temperature of their surroundings may be the reason for the significantly higher purine content of the bee mRNAs. Other stresses could also be involved, e.g., more prone to infections in social insects, as suggested by one of the reviewers of the manuscript.

The HSPs and DNAJs of *A. thaliana* have a higher R content and purines in tracts compared to the

Table 1. Purine content and R- and Y-tract contents of THEG and RSG mRNAs

(1) Species and gene group	(2) REx	(3) *P*	(4) RtrEx	(5) *p*	(6) YtrEx	(7) *p*
A. thaliana						
RPs (61)	50.99	0.041	22.2	$<10^{-6}$	9.1	$<10^{-4}$
HSPs (43)	53.71	10^{-5}	19.8	$<10^{-4}$	11.0	NS
DNAJs (12)	55.09	$<10^{-3}$	23.6	$<10^{-5}$	8.8	0.007
ARSs (35)	52.72	0.029	18.6	0.014	10.5	NS
THEGs (155)	52.48	0.016	20.7	$<10^{-6}$	9.9	$<10^{-3}$
RSGs (199)	**51.76**		**17.2**		**11.0**	
C. elegans						
RPs (67)	52.01	< 0.2	18.9	$<10^{-6}$	10.1	0.095
MtRPs (45)	53.01	0.002	16.9	0.011	9.8	0.048
HSPs (35)	51.63	NS	18.2	$<10^{-6}$	11.9	NS
DNAJs (29)	54.42	$<10^{-4}$	20.6	$<10^{-6}$	10.2	0.158
ARSs (22)	52.42	0.145	17.9	0.002	10.1	NS
THEGs (198)	52.57	$<10^{-3}$	18.4	$<10^{-6}$	10.3	0.059
RSGs (200)	**51.32**		**15.3**		**11.2**	
D. melanogaster						
RPs (57)	50.75	$<10^{-6}$	13.6	$<10^{-3}$	10.2	$<10^{-6}$
MtRPs (17)	52.87	NS	14.6	$<10^{-3}$	7.3	NS
HSPs (15)	53.35	0.028	14.5	$<10^{-3}$	8.0	NS
ARSs (7)	51.88	NS	12.4	NS	8.5	NS
THEGs (96)	51.62	0.008	13.8	$<10^{-4}$	9.2	$<10^{-5}$
RSGs (122)	**52.28**		**12.0**		**7.8**	
A. gambiae						
RPs (11)	52.66	0.117	15.0	$<10^{-3}$	6.5	NS
RSGs (20)	**50.53**		**8.6**		**6.1**	
A. mellifera						
RPs (26)	56.19	0.051	20.6	0.022	6.4	0.11
HSPs (9)	54.59	NS	18.3	NS	8.3	NS
ARSs (4)	55.48	NS	17.0	NS	7.4	NS
THEGs (39)	55.84	0.112	19.8	0.08	6.8	NS
RSGs (54)	**54.62**		**17.5**		**7.6**	
C. intestinalis						
RPs (5)	54.97	0.081	20.8	0.012	8.9	NS
MtRPs (8)	54.11	0.063	16.9	0.092	7.4	NS
HSPs (3)	52.76	NS	16.0	NS	9.3	NS
ARSs (8)	54.08	0.058	17.8	0.063	8.1	NS
THEGs (24)	54.11	0.01	17.9	0.006	8.2	NS
RSGs (30)	**51.97**		**15.0**		**8.9**	
D. rerio						
RPs (38)	53.74	10^{-4}	19.2	$<10^{-4}$	10.0	NS
MtRPs (11)	54.07	0.009	17.3	NS	8.6	0.032
HSPs (13)	54.91	$<10^{-3}$	19.5	$<10^{-3}$	9.1	0.109
DNAJs (12)	53.90	0.007	19.1	$<10^{-3}$	9.8	NS
ARSs (8)	54.08	0.022	17.8	0.005	8.1	0.104
THEGs (82)	54.00	$<10^{-6}$	19.0	$<10^{-6}$	9.6	0.027
RSGs (91)	**51.56**		**15.7**		**10.8**	
T. rubripes						
RPs (5)	58.37	0.003	25.0	0.001	7.2	0.041
ARSs (3)	53.59	NS	19.5	0.12	9.3	NS
THEGs (8)	56.58	0.004	22.9	$<10^{-3}$	8.0	0.039
RSGs (20)	**51.48**		**15.7**		**10.7**	
X. tropicalis						
RPs (7)	54.83	0.187	20.5	0.103	8.3	0.078
THEGs (9)	55.57	0.062	22.9	0.027	8.4	0.055
RSGs (20)	**53.25**		**19.1**		**11.6**	
G. galus						
RPs (23)	54.70	0.039	20.1	0.046	8.7	0.006
HSPs (6)	55.11	0.056	20.4	0.068	9.5	0.099
THEGs (30)	54.75	0.013	20.1	0.018	9.0	0.004
RSGs (40)	**52.85**		**17.6**		**11.6**	
M. musculus						
RPs (50)	53.72	$<10^{-6}$	19.1	$<10^{-4}$	10.5	$<10^{-6}$

(Continued)

Table 1. Continued

(1) Species and gene group	(2) REx	(3) *P*	(4) RtrEx	(5) *p*	(6) YtrEx	(7) *p*
MtRPs (69)	51.82	0.03	16.7	NS	11.7	$<10^{-3}$
HSPs (25)	52.80	0.006	17.7	0.06	11.5	0.003
DNAJs (31)	52.56	0.01	18.6	0.003	12.5	0.174
ARSs (24)	52.77	0.002	17.9	0.027	10.8	$<10^{-4}$
THEGs (199)	52.65	$<10^{-6}$	17.9	$<10^{-4}$	11.4	$<10^{-6}$
RSGs (203)	**51.10**		**16.4**		**13.3**	
H. sapiens						
RPs (54)	53.37	$<10^{-6}$	18.8	$<10^{-5}$	10.7	$<10^{-6}$
MtRPs (59)	52.30	$<10^{-5}$	17.4	0.007	11.8	$<10^{-3}$
HSPs (19)	53.40	10^{-5}	18.9	$<10^{-3}$	11.1	0.001
DNAJs (34)	52.52	$<10^{-4}$	18.8	$<10^{-5}$	12.5	0.023
ARS (21)	52.80	$<10^{-4}$	18.6	$<10^{-3}$	11.2	$<10^{-4}$
THEGs (187)	52.81	$<10^{-6}$	18.4	$<10^{-6}$	11.5	$<10^{-6}$
RSGs (192)	**50.32**		**15.6**		**13.9**	

Note. THEG—targeted groups of genes with a high expression level; RSG—randomly selected genes (control group); REx—the average percentage of purines within the mRNA of a specific gene group; RPs—ribosomal proteins; mtRPs—mitochondrial RPs; HSPs—heat shock proteins; DNAJs—HSPs with a prokaryotic-like HSP40 signature; ARSs—amino-acyl tRNA synthetases; RtrEx—the average percentage of purines in R-tracts (sequential purines with $R_n \geq 5$) in the exons of the group of genes in column 1 (the calculation is done by summation of the purines in R-tracts, divided by the total sum of nucleotides in the exons, and displayed as percentages); YtrEx—the average percentage of pyrimidines in Y-tracts (sequential pyrimidines with $Y_n \geq 5$) in the exons of the group of genes in column 1. *p* values in columns 3, 5, and 7 display the significance of the difference between the REx, RtrEx, and YtrEx of a specific gene group and RSGs (control group; boldface) according to Mann-Whitney *U*-test. In parentheses in the first column is the number of genes analyzed.

Table 2. Comparison of purine content of the mRNAs of THEG genes and RSGs of *Arabidopsis*, invertebrates, and chordate species

Species	OGT/BT	THEG REx	RSG REx
A. thaliana (155, 199)	20°C	52.48	51.76
C. elegans (198, 200)	20°C	52.57	51.32
D. melanogaster (96, 142)	22°C	51.68	52.28
A. gambiae (11, 20)	26°C	52.66	50.53
A. mellifera (39, 54)	32°–36°C	55.75[a]	54.62[a]
Invertebrates (346, 396)		52.68[b]	51.97[c]
C. intestinalis (24, 30)	16°C	54.11	51.93
D. rerio (82, 91)	28°C	54.00	51.96
T. rubrieps (8, 20)	28°C	56.58	51.48
X. tropicalis (9, 20)	23°C	55.57	53.25
G. galus (30, 40)	37°C	54.75	52.85
Nonmammal chordates (153, 211)		54.47[b]	52.13[c]
M. musculus (199, 203)	37°C	52.65	51.10
H. sapiens (187, 192)	37°C	52.81	50.32
Mammals (386, 395)	37°C	52.73[b]	50.72[c]

Note. OGT/BT—optimum growth temperature/body temperature. For other abbreviations, see Note to Table 1. In parentheses in the first column are the numbers of analyzed THEGs (left) and RSGs (right).

[a] *A. mellifera* has a higher purine content within its mRNA compared to the three other invertebrates (the insects and *C. elegans*). The significance of the difference is $p < 10^{-6}$ for both the THEG and the RSG groups (Mann-Whitney *U*-test).

[b] The purine contents of mRNAs of the nonmammalian chordate THEG genes are higher than those of the invertebrate THEG mRNAs ($p < 10^{-6}$ Mann-Whitney *U*-test); there is no significant difference in the purine content of the mammal THEG mRNAs versus the invertebrates ($p > 0.5$).

[c] The purine contents of the mRNAs of the invertebrate and nonmammal chordate RSGs are significantly higher than those of the mammal RSG mRNAs ($p < 10^{-6}$).

control (Table 1, columns 2–5). *A. thaliana* OGT is ~20°C, but the surrounding temperature occasionally reaches 40°C (Hong and Vierling 2000). Therefore, there might be a special need in adaptation of the chaperones to high temperatures (on both the mRNA and the protein levels). The high R and R-tract contents of HSP and DNAJ mRNAs might enable better stability of these mRNAs at high temperatures, hinder the formation of dsRNA (a risk that is elevated at high temperatures), promote (indirectly) the encoding protein stability and enhance their chaperone function (see Paz et al. 2004).

Patterns of Purine and Pyrimidine Composition in Introns

In Table 3 (column 2) we summarize the results of testing our assumption that introns are pyrimidine-biased. The purine content of introns is (on average) less than 50%. This was found in all 12 species, in all THEG subgroups, and in the control groups. Moreover, in the plant and invertebrates, THEG introns generally have lower percentages of purines compared to RSGs (control). This trend weakens toward vertebrate species and disappears, or even becomes replaced by an opposite trend, in chicken and mammals: the introns of the chicken, mouse, and human genes (of both THEG and RSG groups) have the lowest ratio of the percentages of nucleotides in Y-tracts to the percentages of nucleotides in R-tracts compared to all other examined species (Fig. 2). As shown in Table 3 (columns 2 and 3) and summarized in Table 4 (columns 2–4), the average purine content of THEG introns is lower than that of the control group in the tested plant, invertebrate, fish, and frog genes (in total, in 29 of 30 cases). For the overall trend, the deviation from H_0 of no trend is significant ($p < 10^{-6}$ using the Mann-Whitney *U*-test).

Our analysis shows that THEG introns have lower percentages of purines in R-tracts compared to the control genes. In particular, the average percentage of purines in R-tracts in THEG introns was lower than that of the control group in 39 of 42 cases ($p < 10^{-6}$); at the individual level of analysis the difference was significant at $p < 0.05$ in 14 of 42 cases (see Table 3, columns 4 and 5).

Although we did not find a general trend of "compensation" to high deviations from Chargaff's second parity rule, it seems that compensation may exist in species that have relatively large exons and the ratio of exon average length to the introns' average length is between ~0.5 and 2 (Fig. 3).

Purine Bias in "Average" mRNAs in Mammals Is the Lowest Among the Tested Species

As shown in Table 2, column 4, the purine content of mouse and human mRNAs of "average" genes (control group) is significantly lower than that of all of the other 10 eukaryotes examined. The difference is highly significant ($p < 0.00005$ for mouse, and $p < 10^{-6}$ for human). Moreover, the lower R content of the human RSGs compared to those of the mouse is also significant ($p < 0.005$). The relatively low R-bias of the average genes in mammals is reflected in high percentages of pyrimidines in tracts within the mammalian RSG exons compared to the other eukaryotes ($p < 0.003$ for the mouse and $p < 0.0006$ for the human), although no difference between mouse and human was found in this respect ($p > 0.7$). We suggest that the revealed difference between mammals and other species may be related to the elevated levels of splicing control and alternative splicing in mammals. As shown in Fig. 2, the exons of mouse and human RSGs have the lowest ratio of the percentages of nucleotides in R-tracts to the percentages of nucleotides in Y-tracts compared to all other examined species.

The Purine Content of THEG mRNA Seems to Reflect the Major Trends in the Evolution of Splicing

The purine content of THEG mRNAs of *A. thaliana* and invertebrates proved lower than that of the nonmammalian chordate THEG genes ($p < 10^{-6}$ by Mann-Whitney *U*-test for the two comparisons; see Table 2, column 2). We suggest that the difference in purine content is related to the major use of "exon definition" in the splicing process (Robberson et al. 1990; Berget 1995) within the chordate, compared to a very low or moderate use of "exon definition" within *A. thaliana* and invertebrates. What seems to be a contradiction to our last statement is the finding that the purine content of mammals THEG mRNA is not significantly different from that of the invertebrates ($p > 0.5$) but is significantly lower than that of the other chordate species ($p < 10^{-6}$). The lower purine content of the mammalian THEG mRNAs is reflected in higher percentages of pyrimidines organized as pure Y-tracts ($p < 10^{-6}$) compared to other chordates. We suggest that this trend may reflect the elevated levels of splicing control (including alternative splicing) within the mammals. As discussed in the literature, the pyrimidine tracts in the exons are the target sequences of the splicing control system.

We showed previously (Paz et al. 2004) that within prokaryotic thermophiles, the high R-bias and elevated levels of R-tracts of the mRNA are an important evolutionary adaptation to life at high temperatures, and that in the mRNAs of the highly expressed genes, RPs, HSPs, and ARSs, these features are even more pronounced than in average genes. In this study, we showed that in eukarya the R-Y composition and organization of the pre-mRNAs of RPs, HSPs, and ARSs (of both exons and *introns*) are significantly different from that of the average genes. These differences can be attributed to the high level of expression of the three targeted gene groups relative to the control group. The specific R-Y composition of THEG mRNAs enables lower levels of mRNA secondary structures to be formed (structures that slow down the translation rate), higher frequencies of sequences that are targets for proteins that enhance splicing, and lower frequencies of sequences that are targets for proteins that suppress splicing. The differences in R-Y composition of the introns of THEGs and average genes can be attrib-

Table 3. Comparison of purine content and R- and Y-tract contents of the introns of THEGs and RSGs of 12 eukaryotes

(1) Species and gene group	(2) RInt	(3) *p*	(4) RtrInt	(5) *p*	(6) YtrInt	(7) *p*
A. thaliana						
RPs (61)	43.04	0.009	4.8	$<10^{-6}$	17.3	0.05
HSPs (43)	42.87	0.005	7.0	NS	20.7	0.029
DNAJs (12)	43.04	0.101	4.6	0.014	19.8	NS
ARSs (35)	43.50	0.155	6.3	0.164	18.8	NS
THEGs (155)	43.11	$<10^{-3}$	5.7	10^{-6}	18.7	NS
RSGs (199)	**44.15**		**7.0**		**18.5**	
C. elegans						
RPs (67)	45.92	$<10^{-6}$	11.0	$<10^{-6}$	18.4	NS
mtRPs (45)	48.91	NS	16.1	NS	19.1	0.004
HSPs (35)	48.58	NS	13.9	0.112	18.6	0.007
DNAJs (29)	48.30	0.158	14.7	NS	18.0	0.117
ARSs (22)	48.35	0.087	11.8	0.006	16.2	NS
THEGs (198)	47.69	$<10^{-4}$	13.4	$<10^{-4}$	18.3	0.014
RSGs (200)	**49.06**		**14.9**		**17.2**	
D. melanogaster						
RPs (57)	48.35	NS	8.5	$<10^{-3}$	13.1	NS
mtRPs (17)	48.71	NS	12.1	NS	13.9	NS
HSPs (15)	46.94	0.011	9.0	0.045	13.9	NS
ARSs (7)	45.88	0.015	7.5	0.016	15.2	0.101
THEGs (96)	48.02	0.021	9.1	$<10^{-4}$	13.5	NS
RSGs (122)	**48.62**		**10.0**		**12.8**	
A. gambiae						
RPs (11)	47.88	NS	7.9	0.094	12.5	NS
RSGs (20)	**48.73**		**9.4**		**11.7**	
A. mellifera						
RPs (26)	47.68	NS	9.0	NS	13.0	NS
HSPs (9)	48.82	NS	10.7	NS	13.3	NS
ARSs (4)	47.32	NS	8.6	NS	15.9	NS
THEGs (39)	47.98	NS	9.1	NS	13.3	NS
RSGs (54)	**48.79**		**10.4**		**13.1**	
C. intestinalis						
RPs (5)	49.04	NS	7.9	NS	10.0	NS
mtRPs (8)	49.77	NS	8.5	NS	9.2	NS
HSPs (3)	47.58	0.117	7.1	0.079	11.9	NS
ARSs (8)	48.31	0.086	8.4	NS	11.2	NS
THEGs (24)	48.96	NS	8.2	0.057	10.4	NS
RSGs (30)	**49.33**		**9.6**		**10.8**	
D. rerio						
RPs (38)	48.57	0.035	8.9	10^{-5}	11.7	NS
mtRPs (11)	47.54	0.142	8.8	0.002	11.2	NS
HSPs (13)	49.13	NS	10.6	NS	12.5	NS
DNAJs (12)	48.93	NS	9.9	NS	11.5	NS
ARSs (8)	48.31	NS	8.4	NS	11.2	NS
THEGs (82)	48.61	0.121	9.4	$<10^{-4}$	11.7	NS
RSGs (91)	**49.15**		**10.3**		**11.8**	
T. rubripes						
RPs (5)	46.40	NS	6.3	0.03	15.1	NS
ARSs (3)	46.43	NS	6.4	NS	16.0	NS
THEGs (8)	46.41	NS	6.3	0.033	15.5	NS
RSGs (20)	**47.29**		**9.4**		**15.2**	
X. tropicalis						
RPs (7)	48.33	0.215	10.0	0.14	13.5	NS
HSPs (2)	48.21	NS	10.3	NS	14.6	NS
THEGs (9)	48.30	0.194	10.0	0.099	13.8	NS
RSGs (20)	**49.43**		**12.0**		**13.5**	
G. galus						
RPs (23)	49.98	0.026	12.7	0.196	13.6	0.009
HSPs (6)	49.74	0.181	12.7	0.138	13.7	0.016
THEGs (30)	49.88	0.026	12.7	0.077	13.7	0.003
RSGs (40)	**49.09**		**13.4**		**15.6**	

(Continued)

Table 3. Continued

(1) Species and gene group	(2) RInt	(3) *p*	(4) RtrInt	(5) *p*	(6) YtrInt	(7) *p*
M. musculus						
RPs (50)	49.61	0.009	12.6	0.001	13.8	$<10^{-6}$
mtRPs (69)	49.06	0.171	13.5	NS	15.8	0.043
HSPs (25)	49.16	0.133	14.9	NS	15.7	0.038
DNAJs (31)	49.01	NS	13.6	NS	15.5	0.092
ARSs (24)	48.67	NS	13.1	0.084	16.0	NS
THEGs (199)	49.16	0.041	13.4	0.026	15.3	$<10^{-5}$
RSGs (203)	**48.92**		**13.8**		**16.0**	
H. sapiens						
RPs (54)	49.51	0.049	13.2	0.015	14.7	$<10^{-4}$
mtRPs (59)	49.10	NS	13.7	NS	15.5	0.129
HSPs (19)	48.06	NS	12.5	0.052	17.1	NS
DNAJs (34)	48.47	0.08	13.3	0.044	16.4	NS
ARS (21)	48.41	NS	12.5	0.004	16.0	NS
THEGs (187)	48.92	NS	13.2	0.002	15.7	0.022
RSGs (192)	**48.98**		**13.9**		**16.0**	

Note. RInt—the average percentage of purines within the introns of a specific gene group; RtrInt—the average percentage of purines in R-tracts, in the introns of the group of genes in column 1 (the calculation is done by summation of the purines in R-tracts, divided by the total sum of nucleotides in the introns, and displayed as percentages); YtrInt—the average percentage of pyrimidines in Y-tracts, in the introns of the group of genes in column 1. For other abbreviations, see the Note to Table 1. *p* values in columns 3, 5, and 7 display the significance of the difference between the RInt, RtrInt, and YtrInt of specific gene groups and RSGs (control group; boldface) according to Mann-Whitney *U*-test. In parentheses in the first column is the number of genes analyzed.

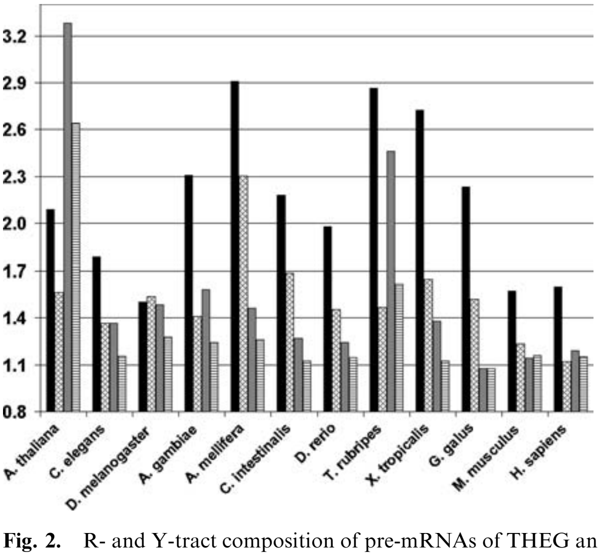

Fig. 2. R- and Y-tract composition of pre-mRNAs of THEG and RSG groups of 12 examined eukaryotic species. THEG, the three targeted highly expressed genes; RSG, random selected genes (control group). Black and hatched columns: the ratio of the percentage of nucleotides in R-tracts to the percentage of nucleotides in Y-tracts in THEG and RSG exons, respectively. Gray and striped columns: the ratio of the percentage of nucleotides in Y-tracts to the percentage of nucleotides in R-tract in THEG and RSG introns, respectively.

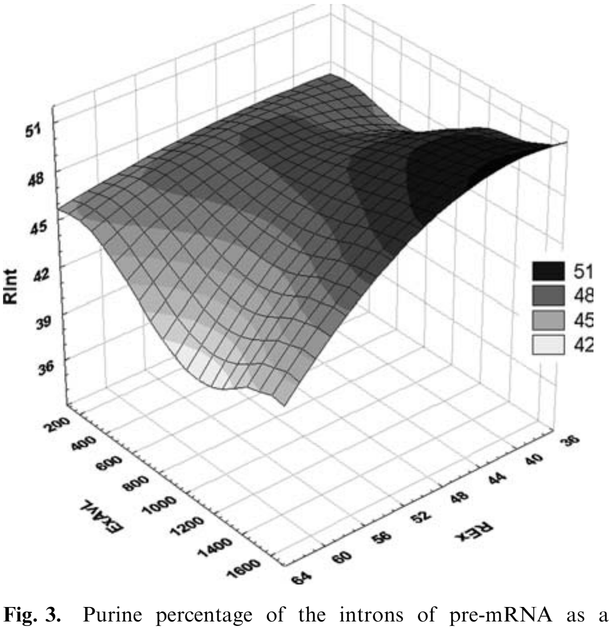

Fig. 3. Purine percentage of the introns of pre-mRNA as a function of the purine percentage of exons and the average exon length. Species included are *A. thaliana, C. elegans, D. melanogaster, A. gambiae, A. mellifera, C. intestinalis, D. rerio, T. rubripes*, and *X. tropicalis.* Rex, purine percentage of mRNA; ExAvL, average length of the exons of the mRNA (number of nucleotides in the mRNA divided by number of exons); RInt, purine percentage of the introns of the pre-mRNA.

uted also to differences in the control of splicing; the introns of THEG genes indeed show lower levels of R-tracts (which might suppress constitutive splicing of the introns). The two major trends in the evolution of splicing, the switch from high use of "intron definition" in the invertebrates to "exon definition" in the vertebrates and the higher use of alternative splicing within the mammals, had influenced the R-Y compositional design of all pre-mRNAs, in general, and the pre-mRNAs of the highly expressed genes, in particular. We suggest that the switch from high use of "intron definition" (in *A. thaliana* and inverte-

Table 4. Comparison of purine content of the introns of THEGs and RSGs of *Arabidopsis*, invertebrates, and chordate species

Species	THEG RInt	RSG RInt	Dif
A. thaliana (155, 199)	43.11	44.21	1.10
C. elegans (198, 200)	47.69	49.06	1.37
D. melanogaster (96, 142)	48.02	48.62	0.60
A. gambiae (11, 20)	47.88	48.73	0.85
A. mellifera (39, 54*)*	47.91	48.79	0.88
Invertebrates (346, 396)	**47.82**[a]	48.87[a]	1.05
C. intestinalis (24, 30)	48.85	49.33	0.48
D. rerio (82, 91)	48.61	49.15	0.52
T. rubrieps (8, 20)	46.41	47.29	0.88
X. tropicalis (9, 20)	48.30	49.43	1.13
Poikilothermic chordates (123, 171)	**48.50**[a]	49.01[a]	0.51
G. galus (30, 40)	49.88	49.09	−0.79
M. musculus (199, 203)	49.16	48.92	−0.24
H. sapiens (187, 192)	48.92	48.98	0.06
Hot-blooded (416, 435)	**49.10**[a]	48.96[a]	−0.14

Note. RInt—the average percentage of purines in the introns; Dif—the difference between the average purine percentage of the introns of RSG and the THEG genes. For other abbreviations, see the Note to Table 1. In parentheses in the first column are the numbers of analyzed THEGs (left) and RSGs (right).

[a]The introns of THEG genes of invertebrates have a significantly lower purine content than those of poikilothermic chordate THEG genes (Mann-Whitney *U*-test, $p = 0.0438$) and hot-blooded vertebrates ($p < 10^{-6}$). There are no differences in the purine content of the introns of RSGs of invertebrates versus those of RSGs of both poikilothermic and hot-blooded vertebrates ($p > 0.75$ and $p > 0.22$, respectively). The introns of THEG genes of the hot-blooded species *G. galus, M. musculus*, and *H. sapiens* have a significantly higher purine content than the poikilothermic *chordate* species ($p = 0.002$). The difference between these species regarding the purine content of the RSG groups is not significant ($p > 0.4$).

brates) to splicing based on "exon definition" (in the vertebrates) is the reason for the elevated levels of purines within the vertebrate THEG mRNAs compared to the former species. In the "second wave" of splicing evolution, the higher eukaryotes evolved elevated rates of alternative splicing. Control and repression of exon splicing might be correlated with an elevated abundance of pyrimidine tracts within exons. Pyrimidine tracts are target sequences for pyrimidine tract binding protein (PTB), a mechanism proposed to be the major contributor to exon splicing silencing (Wagner and Garcia-Blanco 2001). We showed that in mammals, compared to other vertebrates, there is a reduction of purine content and increased level of pyrimidines in tracts (in pre-mRNAs of both THEG and control groups). This might be related to the needs for more flexible splicing control and higher levels of alternative splicing in mammals.

Possible Functions for the Increased Content of Purines and Purine Tracts Within the Exons of THEG Genes

The most obvious explanation for the observed pattern is that THEG proteins need a high abundance of charged amino acids within their sequences compared to "average" proteins. Lysine and glutamic acids are encoded by pure purinic codons. The two other charged amino acids, aspartic acid and arginine, are also encoded by relatively purine-rich triplets (12 of the 18 possible nucleotides of the six arginine codons and 4 of 6 nucleotides of the two codons for aspartic acid). This consideration of the coding potential of R-rich and R-tract-rich mRNAs cannot rule out additional roles for this R-bias. Thus, compositional organization of exon sequences might be adapted to a high efficiency of both splicing and translation. The possibility of the coevolution of codon usage and splicing enhancers was suggested earlier (Schaal and Maniatis 1999). We propose two additional roles for the observed purine bias: (1) high purine bias of mRNAs can minimize the possibility of formation of dsRNA and/or secondary structures, and (2) purine tracts can be used as exonic splicing enhancers (ESEs).

There are a few reasons for the above proposal (2) for R-bias. In vertebrates, splicing is based on "exon definition" (Robberson et al. 1990; Berget 1995; Romfo et al. 2000; Collins and Penny 2006). Presumably, high purine bias and the abundance of pure-purine tracts in the mRNA of THEGs may enhance the splicing rate and efficiency. Purine-rich tracts serve as *cis*-acting sequences for serine/arginine-rich (SR) proteins of the splicing machinery. Some of the well-known ESEs are purine-rich elements. In certain cases these elements are even represented by a consensus sequence $(GAR)_n$ (Liu et al. 1998; Tacke and Manley 1999; Shcaal and Maniatis 1999; Caputi and Zahler 2001; Black 2003; Webb et al. 2005). These sequences are bound by the SR proteins ASF/SF2 and Tra2 of the splicing machinery. It should be mentioned that enhancement of the splicing efficiency is not restricted to alternative spliced pre-mRNA: ESEs might also promote constitutive splicing, and SR proteins are involved in both alternative and constitutive splicing (Bourgeois et al. 2004; Cazalla et al. 2005; Ibrahim et al. 2005). Thus, abundance of the purine tracts might enable an increased rate and efficiency of splicing in these highly expressed genes compared to genes with lower expression.

The Meaning of the Lower Content of Pyrimidine Tracts Within Exons of THEG Genes Compared to the Control Group

The simplest explanation is that the observed pattern resulted mainly from the coding requirements and that the sequences of THEG proteins include fewer hydrophobic amino acids (that are encoded by pyrimidine-biased codons) than average proteins. We suggest that the relatively low frequency of pyrimidine tracts within THEG mRNAs serves two additional roles.

Reducing the formation of secondary structures. This might result from interaction between pyrimidine tracts and the highly abundant purine tracts within THEG mRNA (G·U or A·C non–Watson–Crick pairs can be formed in some circumstances [Meroueh and Chow 1999]). Thus, even if the bases in one tract are not perfectly complementary to the bases in the other tract, undesirable secondary structures might be formed between poly(R) and poly(Y) tracts.

Lowering the rate of THEG exon splicing suppression. In vertebrates, splicing is based mainly on exon definition (Robberson et al. 1990; Berget 1995; Collins and Penny 2006). PTBs binding to pyrimidine tracts in the exon had been suggested to be the main cause of exon splicing silencing (Wagner and Garcia-Blanco 2001), and several models that explain the silencing mechanisms were proposed (Wagner and Garcia-Blanco 2001; Oberstrass et al. 2005; Amir-Ahmady et al. 2005; Ibrahim et al. 2005). It seems reasonable that the frequency of Y-tracts will be lower in the exons of highly expressed genes than in the exons of genes that need more control of the expression on the splicing level.

Possible Functions for High Pyrimidine Bias and Low Abundance of Purine Tracts Within the Introns of THEG Genes in the Plant and Invertebrates

The pyrimidine bias of introns is especially pronounced in THEG genes and coincides with a lower percentage of purine in tracts. We suggest that this composition of THEG gene introns is an adaptation to high expression, in two of the following (not mutually exclusive) aspects.

(1) The foregoing specific intron composition might lower the chance of formation of RNA secondary structures by undesirable bonds between the intronic polypyrimidine tract located near the 3' splice site, an essential *cis*-acting component of splicing (Ruskin and Green 1985; Roscigno et al. 1993; Coolidge et al. 1997), and the polypurine tract. This should be more important in species where splicing of (many or most) pre-mRNAs is based on "intron definition" (in invertebrates and *A. thaliana*, respectively).

(2) Many ESEs, the target sequences for the binding of SR proteins in exons, are purine-rich (Black 2003; Webb et al. 2005). There are reports that SR proteins can suppress splicing when bound to sequences located within introns (Kanopka et al. 1996; Dauksaite and Akusjarvi 2002; Ibrahim et al. 2005). In addition, sequences that are the binding targets for SR proteins are ESEs when located in exons, but when inserted to introns, they can cause inactivation of a 3′ splice site located downstream (Ibrahim et al. 2005). Therefore Maniatis and coworkers (Ibrahim et al. 2005) proposed that in constitutively spliced introns, SR protein binding sites should be rare. It is reasonable that in species with lower rates of alternative splicing, the abundance of sequences that are used as ESEs will be low in introns, especially in the introns of THEG genes.

The chicken and mammals have the lowest pyrimidine bias within THEG introns compared to all the other species examined. The elevated frequency of purines in the introns of THEG genes of these higher eukaryotes might enable higher rates of alternative splicing, by the inclusion (in certain cases) of parts of the introns in the mRNA. As the composition of all mRNAs is, on the average, R-biased, and THEG genes have an even higher R-bias than the average, intronic sequences with higher purine content have (in general) a higher probability of becoming a part of mature mRNAs, especially in THEG genes. For such events to occur, it is also required that other *cis*-components for the splicing machinery will be properly organized.

Final Remarks and Conclusions

We suggest that the R-Y composition of pre-mRNAs had influenced the evolution of splicing, and vice versa, was affected by the pressures resulting from evolutionary changes in the splicing machinery. As mRNAs tend to be, on the average, purine-biased, distinguishing between exonic and intronic sequences might be easier if intronic sequences become pyrimidine-biased. And indeed, the polypyrimidine tract located downstream from the branch point, near the 3′ splice site, is an essential *cis*-component of splicing. The pyrimidine bias of the introns, which might enable a high efficiency of splicing, is conserved in species with a high use of "intron definition" in the splicing process (*A. thaliana* and invertebrates). In these species, the introns of highly expressed THEG genes seem to be "superintrons," better adapted to proper recognition by the splicing machinery components without unnecessary disturbances. Recently, Andolfatto (2005) adopted the McDonald–Kreitman test (1991) for noncoding sequences and estimated that positive selection affected ~20% nucleotides in introns. If the constraint intronic nucleotides (estimated relative to fourfold synonymous sites) are also considered, then more than 50% of the nucleotides in the introns are considered to be functionally relevant.

The switch from high use of "intron definition" in the splicing process in invertebrates to "exon definition" in the vertebrates (due to the increasing intron length) had influenced the R-Y compositional design of all pre-mRNAs, in general, and the pre-mRNAs of the highly expressed genes, in particular. The mRNAs of highly expressed genes should have a higher abundance of ESEs and a lower level of exonic splicing silencers, despite the constraints of the coding role. It seems very reasonable that purine-biased sequences were chosen to be *cis*-acting ESEs (binding targets of the SR proteins), due to their high abundance in the mRNAs of RPs, HSPs, and ARSs (proteins that are all highly expressed due to their key role in the cell metabolism). There could also have been a need to lower the amount of target sequences for the binding of SR proteins in the introns, and this trend might be even more important for THEG genes. We believe that the pre-mRNA R-Y composition of both exons and introns, especially the distribution of homotracts, is an important layer of gene organization, which strongly influences the expression level of eukaryotic genes.

Acknowledgments. We thank Ed Trifonov, Zakharia Frenkel, Axel Meyer, and anonymous reviewers for their helpful comments and suggestions. This work was supported by the Israeli Ministry of Absorption and by the Ancell-Teicher Research Foundation for Molecular Genetics and Evolution. A.P. was supported by a scholarship in bioinformatics from the Eshkol Foundation of the Israeli Ministry of Science and Technology.

References

Amir-Ahmady B, Boutz PL, Markovtsov V, Phillips ML, Black DL (2005) Exon repression by polypyrimidine tract binding protein. RNA 11:699–716

Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100:3889–3894

Bec G, Kerjan P, Waller JP (1994) Reconstitution in vitro of the valyl-tRNA synthetase-elongation factor (EF) 1 beta gamma delta complex. Essential roles of the NH2-terminal extension of valyl-tRNA synthetase and of the EF-1 delta subunit in complex formation. J Biol Chem 269:2086–2092

Bell SJ, Forsdyke DR (1999) Deviations from Chargaff's second parity rule correlate with direction of transcription. J Theor Biol 197:63–76

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72:291–336

Bourgeois CF, Lejeune F, Stevenin J (2004) Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog Nucleic Acid Res Mol Biol 78:37–88

Caplan AJ, Cyr DM, Douglas MG (1993) Eukaryotic homologues of *Escherichia coli* dnaJ: a diverse protein family that functions with hsp70 stress proteins. Mol Biol Cell 4:555–563

Caputi M, Zahler AM (2001) Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem 276: 43850–43859

Cazalla D, Newton K, Caceres JF (2005) A novel SR-related protein is required for the second step of Pre-mRNA splicing. Mol Cell Biol 25:2969–2980

Collins L, Penny D (2006) Investigating the intron recognition mechanism in eukaryotes. Mol Biol Evol 23:901–910

Coolidge CJ, Seely RJ, Patton JG (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Res 25:888–896

Dauksaite V, Akusjarvi G (2002) Human splicing factor ASF/SF2 encodes for a repressor domain required for its inhibitory activity on pre-mRNA splicing. J Biol Chem 277:12579–12586

Fewell SW, Travers KJ, Weissman JS, Brodsky JL (2001) The action of molecular chaperones in the early secretory pathway. Annu Rev Genet 35:149–191

Forsdyke DR (1999) Heat shock proteins as mediators of aggregation-induced 'danger' signals: implications of the slow evolutionary fine-tuning of sequences for the antigenicity of cancer cells. Cell Stress Chaperones 4:205–210

Frydman J (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. Annu Rev Biochem 70:603–647

Fujimori S, Washio T, Tomita M (2005) GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 6:26

Hartl FU, Hayer-Hartl M (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. Science 295:1852–1858

Herruer MH, Mager WH, Woudt LP, Nieuwint RT, Wassenaar GM, Groeneveld P, Planta RJ (1987) Transcriptional control of yeast ribosomal protein synthesis during carbon-source upshift. Nucleic Acids Res 15:10133–10144

Hong SW, Vierling E (2000) Mutants of *Arabidopsis thaliana* defective in the acquisition of tolerance to high temperature stress. Proc Natl Acad Sci USA 97:4392–4397

Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR (2001) A compendium of gene expression in normal human tissues. Physiol Genomics 7:97–104

Ibrahim el C, Schaal TD, Hertel KJ, Reed R, Maniatis T (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. Proc Natl Acad Sci USA 102:5002–5007

Kanopka A, Muhlemann O, Akusjarvi G (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. Nature 381:535–538

Karkas JD, Rudner R, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl Acad Sci USA 60:915–920

Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361:13–37

Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res 10:228–236

Lee SW, Cho BH, Park SG, Kim S (2004) Aminoacyl-tRNA synthetase complexes: beyond translation. J Cell Sci 117:3725–3734

Le Hir H, Moore MJ, Maquat LE (2000) Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev 14:1098–1108

Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12:1998–2012

Lobry JR, Chessel D (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J Appl Genet 44:235–261

Marygold SJ, Coelho CM, Leevers SJ (2005) Genetic analysis of RpL38 and RpL5, two minute genes located in the centric heterochromatin of chromosome 2 of *Drosophila melanogaster*. Genetics 169:683–695

McDonald J, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654

Meroueh M, Chow CS (1999) Thermodynamics of RNA hairpins containing single internal mismatches. Nucleic Acids Res 27:1118–1125

Meyuhas O (2000) Synthesis of the translational apparatus is regulated at the translational level. Eur J Biochem 267:6321–6330

Negrutskii BS, Shalak VF, Kerjan P, El'skaya AV, Mirande M (1999) Functional interaction of mammalian valyl-tRNA synthetase with elongation factor EF-1alpha in the complex with EF-1H. J Biol Chem 274:4545–4550

Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, Allain FH (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science 309:2054–2057

Park SG, Ewalt KL, Kim S (2005) Functional expansion of aminoacyl-tRNA synthetases and their interacting factors: new perspectives on housekeepers. Trends Biochem Sci 30: 569–574

Paz A, Mester D, Baca I, Nevo E, Korol A (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. Proc Natl Acad Sci USA 101:2951–2956

Paz A, Kirzhner V, Nevo E, Korol A (2005) Coevolution of DNA-interacting proteins and genome "dialect". Mol Biol Evol 23:56–64

Perry RP (2005) The architecture of mammalian ribosomal protein promoters. BMC Evol Biol 5:15

Prabhu VV (1993) Symmetry observations in long nucleotide sequences. Nucleic Acids Res 21:2797–2800

Robberson BL, Cote GJ, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84–94

Romfo CM, Alvarez CJ, van Heeckeren WJ , Webb CJ, Wise JA (2000) Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. Mol Cell Biol 20:7955–7970

Roscigno RF, Weiner M, Garcia-Blanco MA (1993) A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing. J Biol Chem 268:11222–11229

Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA 60:921–922

Ruskin B, Green MR (1985) Role of the 3′ splice site consensus sequence in mammalian pre-mRNA splicing. Nature 317:732–734

Sang Lee J, Gyu Park S, Park H, Seol W, Lee S, Kim S (2002) Interaction network of human aminoacyl-tRNA synthetases and subunits of elongation factor 1 complex. Biochem Biophys Res Commun 291:158–164

Schaal TD, Maniatis T (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol Cell Biol 19:261–273

Seshaiah P, Andrew DJ (1999) WRS-85D: A tryptophanyl-tRNA synthetase expressed to high levels in the developing *Drosophila* salivary gland. Mol Biol Cell 10:1595–1608

Smithies O, Engels WR, Devereux JR, Slightom JL, Shen S (1981) Base substitutions, length differences and DNA strand asymmetries in the human G gamma and A gamma fetal globin gene region. Cell 26:345–353

Szybalski W, Kubinski H, Sheldrick P (1966) Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. Cold Spring Harb Symp Quant Biol 31:123–127

Tacke R, Manley JL (1999) Determinants of SR protein specificity. Curr Opin Cell Biol 11:358–362

Tatarinova T, Brover V, Troukhan M, Alexandrov N (2003) Skew in CG content near the transcription start site in *Arabidopsis thaliana*. Bioinformatics 19(Suppl 1):i313–i314

Tautz J, Maier S, Groh C, Rossler W, Brockmann A (2003) Behavioral performance in adult honey bees is influenced by the temperature experienced during their pupal development. Proc Natl Acad Sci USA 100:7343–7347

van der Velden AW, van Nierop K, Voorma HO, Thomas AA (2002) Ribosomal scanning on the highly structured insulin-like growth factor II-leader 1. Int J Biochem Cell Biol 34: 286–297

van Ruissen F, Jansen BJ, de Jongh GJ, Zeeuwen PL, Schalkwijk J (2002) A partial transcriptome of human epidermis. Genomics 79:671–678

Wagner EJ, Garcia-Blanco MA (2001) Polypyrimidine tract binding protein antagonizes exon definition. Mol Cell Biol 21:3281–3288

Warner JR (1999) The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 24:437–440

Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol Genomics 2:143–147

Webb CJ, Romfo CM, van Heeckeren WJ, Wise JA (2005) Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. Genes Dev 19:242–254

Yu Y, Zhang C, Zhou G, Wu S, Qu X, Wei H, Xing G, Dong C, Zhai Y, Wan J, Ouyang S, Li L, Zhang S, Zhou K, Zhang Y, Wu C, He F (2001) Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. Genome Res 11:1392–1403

Zuckerkandl E (1986) Polite DNA: functional density and functional compatibility in genomes. J Mol Evol 24:12–27