ISSN 0026-8933, Molecular Biology, 2008, Vol. 42, No. 1, pp. 133–145. © Pleiades Publishing, Inc., 2008.
Original Russian Text © O.M. Korzinov, T.V. Astakhova, P.K. Vlasov, M.A. Roytberg, 2008, published in Molekulyarnaya Biologiya, 2008, Vol. 42, No. 1, pp. 150–162.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 577.21;579.23"315

Statistical Analysis of DNA Sequences in the Neighborhood of Splice Sites

O. M. Korzinov[a], T. V. Astakhova[b, c], P. K. Vlasov[d], and M. A. Roytberg[b, c]

[a] *Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow Region, 141700 Russia; e-mail: korzinov@gmail.com*
[b] *Institute of the Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*
[c] *Pushchino State University, Pushchino, Moscow Region, 142290 Russia*
[d] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia; e-mail: vlasov@imb.ac.ru*

Received March 22, 2007

Accepted for publication May 16, 2007

Abstract—Prediction of gene sequences and their exon–intron structure in large eukaryotic genomic sequences is one of the central problems of mathematical biology. Solving this problem involves, in particular, high-accuracy splice site recognition. Using statistical analysis of a splice site-containing human gene fragment database, some characteristic features were described for nucleotide sequences in the splicing site neighborhood, the frequencies of all nucleotides and dinucleotides were determined, and those with frequencies increased or decreased in comparison to a random sequence were identified. The results can be used in sequence annotation, splicing site prediction, and the recognition of the gene exon–intron structure.

DOI: 10.1134/S0026893308010202

Key words: splice sites, exon–intron structure of a gene, statistical sequence analysis

INTRODUCTION

Analysis of an enormous number of the currently available genomic sequences depends on the development and refinement of annotation instruments, both effective and accurate, in particular, to make out the exon–intron structure of genes. In spite of a large number of programs and publications, the problem of recognizing the exon–intron boundaries still exists and has a great theoretical and applied importance [1]. A substantial body of information has been accumulated concerning the diversity of splice site consensus sequences [2, 3] and the features of coding and noncoding regions; however, it remains an extremely intricate problem to determine the exon–intron boundaries with absolute accuracy.

The first studies on bioinformatic splice site analysis date back to the 1980s [4]. Using a relatively small data sample then available, it was shown that the great majority of the splice sites contained conserved dinucleotides, but the number of exceptions was also considerable (see [3] for the last classification of canonical and noncanonical splice sites). Position-specific weight matrices (PSWM) were built for preferred nucleotides at the positions next to the conserved nucleotides; in particular, a preference was shown for T prior to an acceptor site, as well as for GT after AG at the acceptor site and AG before GT at the donor site [4]. These results were later confirmed with much larger data sets, and variations in PSWM were studied in different species and taxa [5].

All current studies on splice site diversity statistics can be conventionally categorized as belonging to a basic or a recognition field. Basic research aims at investigating the physicochemical mechanisms of splicing [6], evolution of splicing [7], effects of splice site mutations on various diseases [8], etc. A comprehensive review summarizes the works in this field [1]. The use of the methods of comparative genomics is also worth mentioning [7].

Research of the second field concerns the development of mathematical methods for recognizing the exon–intron boundaries. This field includes, for example, studies on the use of PSWM [9]. The recent development of these methods has been based on hidden Markov models (HMMs) [10], neuron networks [11], or Bayesian sensors [12].

An essential problem is that a visual interpretation is nearly impossible for the above methods of splice site recognition (except for the "plain" frequency profiles). Our purpose was to overcome this drawback and to provide a more illustrative presentation of statistical results (nucleotide and dinucleotide frequen-

cies, etc.), which are typically hidden in recognition program parameters. In addition, our study was inspired by novel developments of specialized databases that enable analysis of unprecedented amounts of information (see [13] and http://www.oxfordjournals.org/nar/database/subcat/1/3 for a review of the available databases). Detailed research of dinucleotide frequencies has already been undertaken. However, only the frequencies of conserved dinucleotides AG and GT have been considered [14–16], and no currently existing method can recognize splice sites with absolute fidelity. Solving this problem will probably require deep insight into the splicing mechanism.

EXPERIMENTAL

Splice site sample. The study was performed using the EDAS database [17]. To compile the database, genomic sequences were aligned to protein, mRNA, and EST sequences. Thus, all splice sites investigated can be considered experimentally verified. Of the 8234 human exon–intron genomic fragments contained in the database, we extracted donor and acceptor splice sites (conserved dinucleotides) and their flanking sequences. At the next step, all splice sites with noncanonical sequences (i.e., non-GT donor sites and non-AG acceptor sites) were removed from the sample set. The length of the flanking sequences was set as 40 nt (that is, each fragment was 82 nt), and the sequences were brought to the following form convenient for automated analysis:

>*gene_440545* (gene ID in the EntrezGene database) *actgtctcgctggctgcagcgtgtggtccccttaccagagGTaaagaagagatggatctccactcatgttgtag-aca*,

where the donor site is at positions 40–41;

tttctctctctctttctctcctctgtcttttcccacaAGtcgaggatgcgaagagaaggtggctgtctgcaaaccag,

where the acceptor site is at positions 40–41. Positions are numbered from the 5' sequence end.

The resulting sample set of splice sites contained 192 557 donor and acceptor sites and their neighborhoods.

Characteristics of splice sites and their neighborhood. Analyzing the features of nucleotide sequences in the splice site neighborhoods, we calculated the totals for every nucleotide and every dinucleotide at each position and conditional frequencies of dinucleotides with respect to nucleotides, i.e., the ratio of the dinucleotide number to the product of the numbers of the component nucleotides.

The exon and intron neighborhoods of splice sites were analyzed separately.

RESULTS

Preliminary Notes

(1) The conserved dinucleotide of a splice site always occupies positions 0 and +1 in the sequences under analysis. For an acceptor site, the exon lies to the right, and the intron to the left; for a donor site, it is the other way round.

(2) The phrase "fragment S is located at position k" means that the first nucleotide of S is at position k, the second one is at position k + 1, etc.

Principal Results

Acceptor sites. We determined the nucleotide frequencies at each position of the acceptor site neighborhood. The results are shown in Fig. 1a; data for positions [–2, 6] are given to a larger scale in Fig. 1b.

Our data confirm the already known facts with a sample set that is considerably larger than those used in earlier studies. In the intron part, a considerable excess of Ts and, to a lesser extent, Cs and a correspondent deficit of As and, to a lesser extent, Gs was observed from position –3 outwards. In the exon part, the donor site "shadow" (GT) was visible at positions 2 and 3. The periodicity observed from position 4 onwards has been interpreted elsewhere [18] as a consequence of the nucleotide composition of genetic code triplets, with a 3-nt period. Positions –2, –1, 3, and 4 were characterized by unique nucleotide frequency distributions (see [4]). Our study provides more accurate additions to these data.

For example, in the intron region, three zones can be distinguished: [–3, –8] (*a*), [–9, –20] (*b*), and [–21, –40] (*c*). At position –8, the frequency of T had its local maximum of 0.53, while the frequencies of the other nucleotides had local minimums. Within the zone [–3, –8], a local minimum of the T frequency was at position –4 and equaled to 0.45; thus, the frequency variation amounted to about 15%. Starting from position –9, all frequencies changed monotonously: the T and C frequencies fell, while the A and G frequencies grew. The frequencies of T and A changed considerably faster than those of C and G. At position –20, the frequencies of A and C became equal. In region [–21, –40], the A and C frequencies changed little and the C frequency very slowly decreased. The T frequency reached 0.31 at position –30 and further decreased very slowly, reaching 0.3 at position –40. The frequency of G grew nearly uniformly in region [–21, –40] and reached 0.23 at position –40.

It is worth mentioning that, in the exon neighborhood, the cytosine number varied between the phases of the 3-nt period with a much smaller amplitude than the numbers of the other three nucleotides and its periodicity itself was less pronounced, especially close to the splice site (in region [6, 40], the autocorrelation

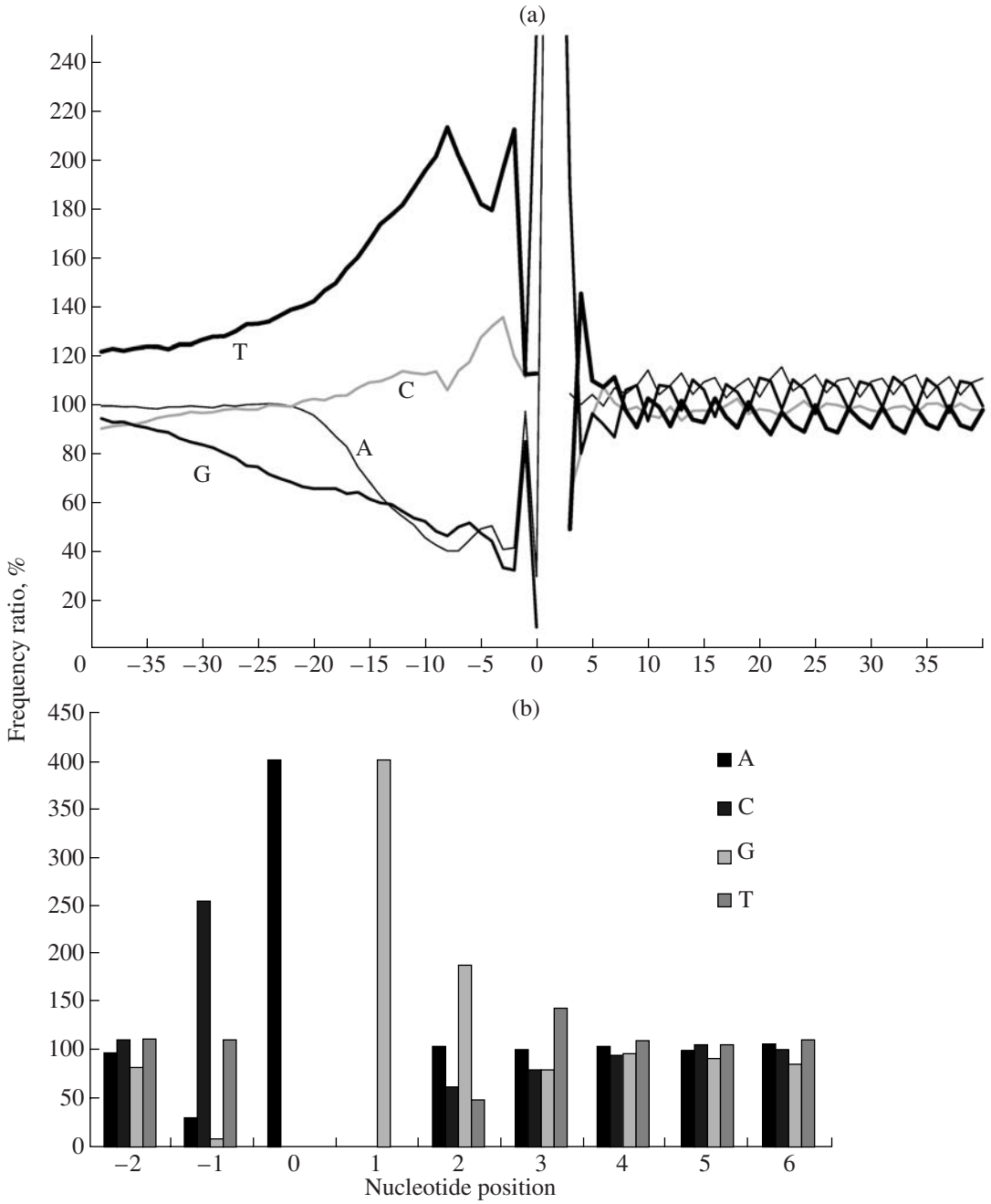

Fig. 1. Nucleotide frequencies in the intron and exon neighborhoods of acceptor sites. (a) Complete neighborhood (positions [–40, +38]). X axis, nucleotide positions within the intron (–1 to –40 from right to left) and the exon (2 to 38 from left to right). Y axis, the ratio of the observed frequency to the frequency expected for a random sequence. The conserved dinucleotide of the acceptor splice site (AG) occupies positions 0 and 1 (the natural frequency peaks at these positions are not shown). (b) Nucleotide frequencies in the [–2, 6] acceptor site neighborhood. Block heights correspond to the ratio of the observed frequency to the frequency expected for a random sequence (taken equal to 0.25, assuming equal probabilities of the four nucleotides).

coefficients for a 3-nt shift were about 0.8 for A, T, and G and about 0.3 for C). For the other nucleotides, the periodicity, strictly speaking, began at position 6 and the frequencies at positions 3, 4, and 5 were transitional. The complete data on nucleotide frequencies for different phases are available at http://www.imb.ac.ru/splicingsites/data.zip. Figure 1b shows, to a larger scale, the nucleotide frequency distribution at transitional positions [–2, 6]; in particular, the shadow of the GT donor splice site at positions 2–3 is seen [5].

At the next step, we analyzed the dinucleotide characteristics. The percent ratios of the observed frequencies of each dinucleotide to the expected average

Table 1. Dinucleotide frequencies in the [–40, –3] and [3, 40] acceptor site neighborhoods

Dinucleotide	Intron	Exon
AA	86	123
AC	68	86
AG	55	127
AT	102	91
CA	85	122
CC	137	112
CG	21	42
CT	170	112
GA	57	125
GC	66	98
GG	60	108
GT	83	78
TA	82	56
TC	145	93
TG	129	135
TT	254	94

Note: Here and in Table 3, data are given as the percent ratio of the observed frequency to the expected random frequency (1/16). The grouping of dinucleotides with similar relative frequencies is shown by color.

frequency of 1/16 in the region of interest are listed in Table 1.

Qualitatively, dinucleotides can be grouped depending on their frequencies (separately for the exon and intron regions). In general, for dinucleotides with similar overall frequencies, the frequencies depended on the position in the same fashion. Thus, to avoid overcharging, Fig. 2 shows only the curves for a few model dinucleotides, each representing a group with similar overall frequencies. The complete data are available at http://www.imb.ac.ru/splicingsites/data.zip.

In the intron region, dinucleotide frequencies correlated with increased frequencies of T and C and decreased frequencies of A and G. However, dinucleotide frequencies cannot be reduced to nucleotide frequencies. In particular, Fig. 2 shows, a deficit of CG and, to a lesser extent, of TA; this "disuse" (not restricted to the splice site neighborhoods) has previously been demonstrated for vertebrate [19] and plant [20] genomes. In addition, the AG number dropped in [–25, –1]. A drastic fall in GA, GG, and AA abundance and a decrease in GT frequency in [–6, –1] are also worth mentioning. In addition, the decrease in the AG frequency in [–25, –1] correlates with a decreased adenine frequency in this region (Fig. 1), but cannot be thus explained completely.

To investigate the dinucleotide selection, we computed the relative (conditional) dinucleotide frequencies for different positions, using the following equation:

$$PCond_n(XY) = P_n(XY)/(P_n(X)P_{n+1}(Y)), \qquad (*)$$

where $P_k(X)$ is the frequency of nucleotide X at position k and $P_k(XY)$ is the frequency of dinucleotide XY at position k. The equation refers to the region to the

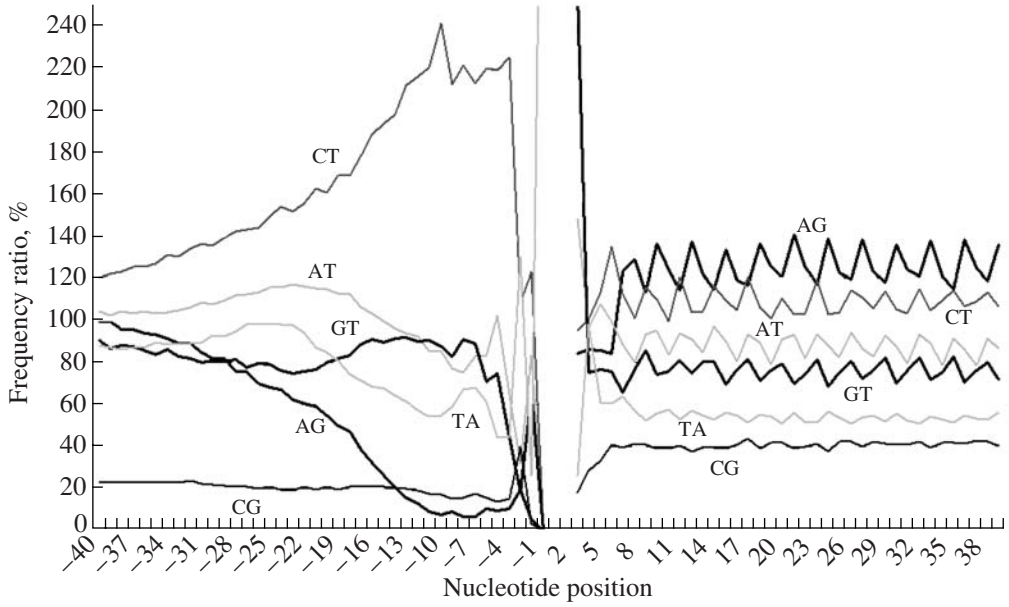

Fig. 2. Frequencies of some dinucleotides in the acceptor site neighborhood. The axes are as in Fig. 1.

right of the conserved dinucleotide; for the regions to the left, $P_{n-1}(Y)$ should be substituted with $P_{n+1}(Y)$.

Conditional dinucleotide frequencies in the intron and exon neighborhoods are characterized in Table 2 and graphically presented in Fig. 3. As shown, the most pronounced deviations of the mean from the expected conditional frequency of 1 were observed in the intron region for AA, GG, and TG (higher frequencies) and for CG, TA, and AG (lower frequencies). As shown in Fig. 3b, a decreased AG content is compensated for by an increase in the contents of AA, GG, and TG to achieve the average total. In the exon region, the deviations of conditional frequencies from unity were less pronounced, except for decreased CG and TA frequencies (as in the total genome) and somewhat increased contents of TG and, to a lesser extent, GA, CT, CA, CC, and AG. Note that, unlike in the intron region, there is no selection against AG in the exon. The conditional frequencies of the above dinucleotides are increased to compensate for CG and TA deficiency. Relatively high TG frequencies can probably be explained by the usage of various amino acids and their codons. In general, the conditional frequency variance averaged over all dinucleotides in the intron was considerably higher than in the exon region (0.11 and 0.04, respectively).

Donor sites. Figure 4a shows the nucleotide frequencies at each position. The transitional zone characterized by obvious preferences for particular nucleotides at individual positions includes five positions of the intron, [2, 6], and five positions of the exon, [–5, –1]; deviations from the stationary sequence were most pronounced in regions [–3, –1] and [2, 5]. Figure 4b shows the nucleotide frequency distribution in the transitional zone. These data agree with the results obtained previously with a smaller sample [4]. In particular, a shadow of the acceptor site AG is clearly visible at exon positions –2 and –1.

In general, nucleotide frequencies in the exon neighborhood of the donor site behaved in the same way as in the exon neighborhood of the acceptor site, including a less pronounced periodicity for cytosine. The average nucleotide frequencies in the exon neighborhoods of the acceptor and donor sites differed by about 5.5% for cytosine and about 2–3% for the other nucleotides. In the stationary part of the intron region, we observed a considerable preference for T (13% above the 25% expected for a uniform nucleotide distribution) along with A and C deficiency (10 and 7% on average, respectively). The frequency of C slightly decreased and the frequency of A slightly increased with increasing distance from the splice site. The portion of G was close to the expected 25%. Unlike with the acceptor site, individual nucleotide frequencies varied insignificantly among different positions.

Next, we discuss the frequencies of dinucleotides in the donor site neighborhood. Table 3 lists the percent ratios of the observed frequency of each dinucleotide to the expected statistical average of 1/16 in the intronic and exonic donor site neighborhoods (outside the transitional zone [–5, 6]). Based on these data, dinucleotides can be conventionally divided in a few

Table 2. Minimum, maximum, mean, and variance for the conditional frequencies of individual dinucleotides in the intron and exon acceptor site neighborhoods

Exon				
dinucleotide	minimum	maximum	mean	variance
AA	1.04	1.26	1.09	0.04
AC	0.81	0.89	0.83	0.02
AG	0.94	1.20	1.15	0.06
AT	0.88	1.02	0.91	0.03
CA	1.00	1.25	1.17	0.05
CC	1.14	1.43	1.20	0.05
CG	0.37	0.47	0.42	0.03
CT	1.10	1.28	1.23	0.03
GA	1.09	1.23	1.15	0.03
GC	0.92	1.07	0.99	0.03
GG	0.91	1.19	1.02	0.04
GT	0.78	0.97	0.82	0.03
TA	0.53	0.68	0.56	0.02
TC	0.87	1.07	1.01	0.04
TG	1.34	1.79	1.40	0.07
TT	0.94	1.14	1.06	0.04
Intron				
dinucleotide	minimum	maximum	mean	variance
AA	1.20	2.40	1.47	0.33
AC	0.71	1.05	0.87	0.07
AG	0.35	1.14	0.88	0.26
AT	0.73	1.10	0.92	0.07
CA	0.83	1.28	1.08	0.06
CC	1.02	1.49	1.26	0.08
CG	0.29	0.44	0.32	0.03
CT	0.81	1.17	1.08	0.09
GA	0.90	1.67	1.07	0.13
GC	0.77	1.12	0.98	0.06
GG	1.14	2.37	1.35	0.24
GT	0.55	1.01	0.84	0.08
TA	0.59	0.85	0.73	0.05
TC	0.68	0.99	0.90	0.04
TG	1.19	1.45	1.34	0.08
TT	0.95	1.36	1.08	0.07

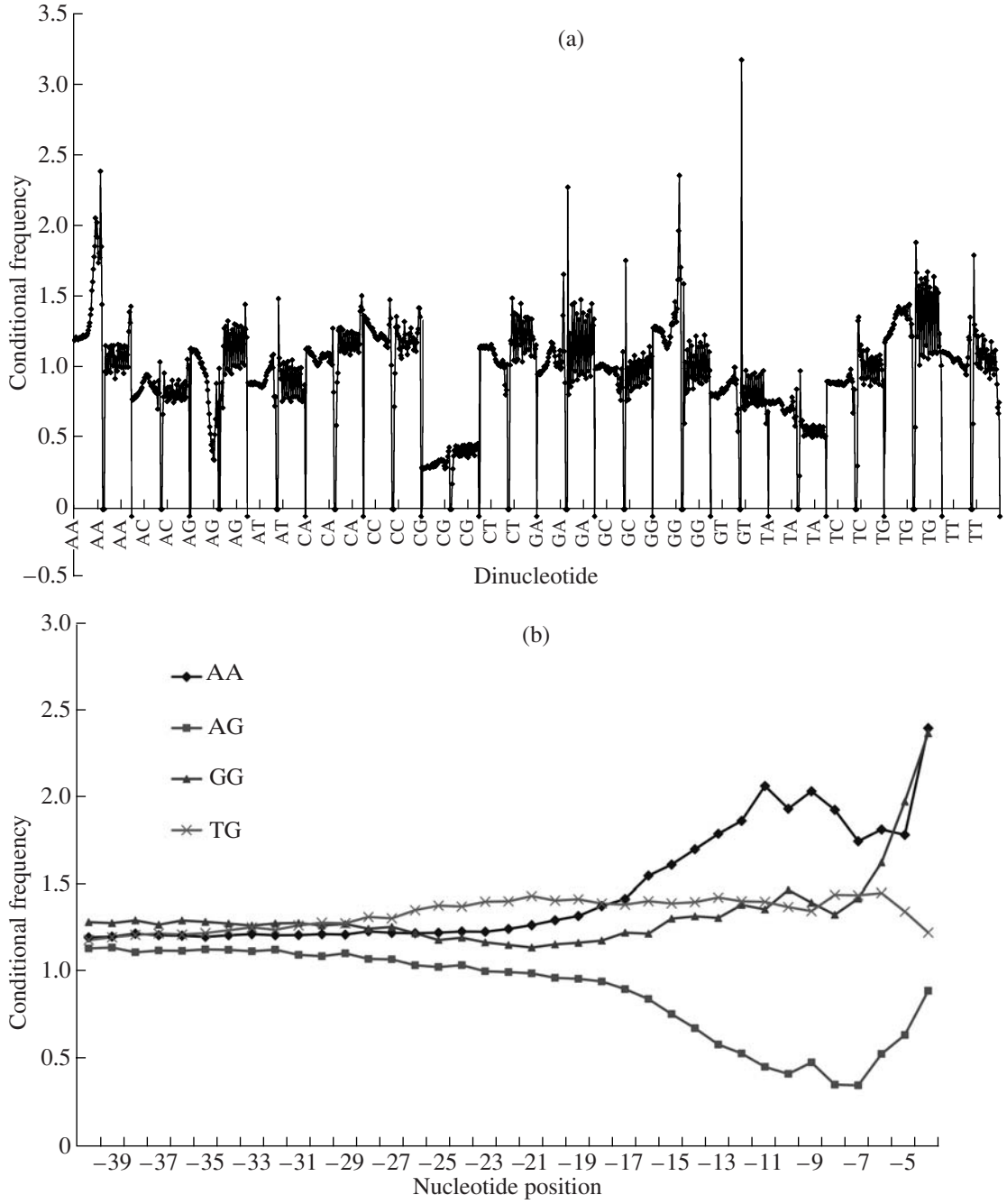

Fig. 3. Conditional dinucleotide frequencies in the "distant" acceptor site neighborhood. (a) Positions [–40, –4] and [4, 38]. X axis, dinucleotides in the alphabetic order. Y axis, the ratio of the dinucleotide frequency to the frequencies of the component nucleotides. For the sake of convenience, the chart is separated with vertical lines corresponding to positions 0 (marked on the X axis) and 40 (marked below the X axis) for each dinucleotide. (b) Positions [–40, –4]. X axis, positions; Y axis, conditional frequencies for AA, AG, GG, and TG.

groups (shown in different colors) according to their frequencies.

The dinucleotide frequencies in the donor site neighborhood are presented in Figs. 5 and 6. The dinucleotide frequencies normalized to the expected frequency of 1/16 of the total sample size are shown in Fig. 5. Each of the Figs. 5a and 5b comprises 2 × 16 charts separated with vertical lines. Figure 6 presents the conditional dinucleotide frequencies related to the product of the component nucleotide frequencies.

As it follows from Fig. 5a, the least frequent dinucleotides outside zone [–5, 6] are CG, TA, and AC in

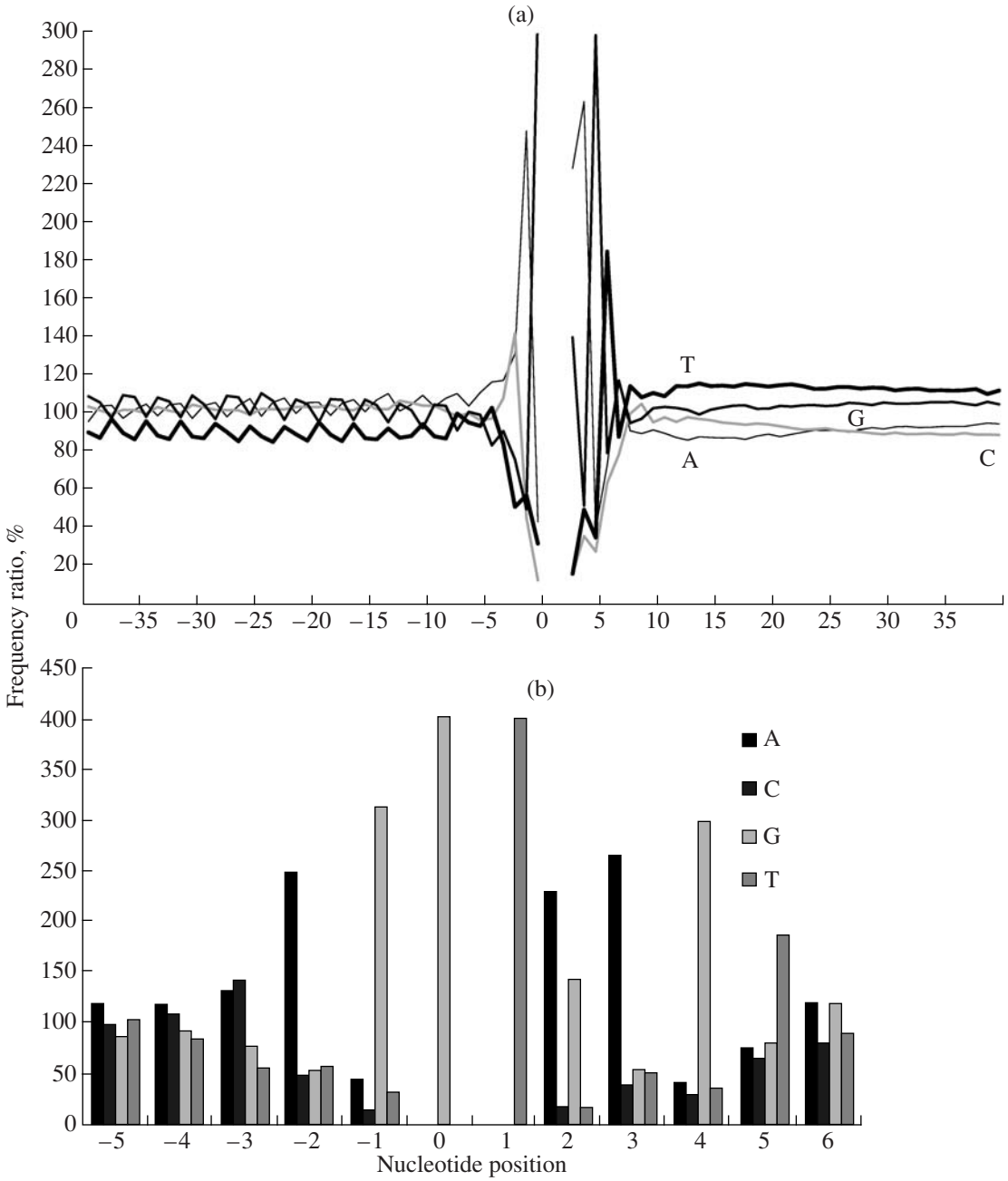

Fig. 4. Nucleotide frequencies in the intron and exon neighborhood of donor sites. (a) Complete neighborhood. X axis, nucleotide positions within the intron (2 to 38 from left to right) and the exon (–1 to –40 from right to left). Y axis, the ratio of the observed frequency to the frequency expected for a random sequence. The conserved dinucleotide of the donor site (GT) occupies positions 0 and 1 (natural frequency peaks at these positions are not shown). (b) Nucleotide frequencies in the [–5, 6] donor site neighborhood. Block heights correspond to the ratio of the observed frequency to the frequency expected for a random sequence.

the intron and GT in the exon. CG and TA are avoided for the same reasons as in the neighborhood of acceptor sites (discussed above). The deficit of GT, as well as the AG deficit nearby acceptor sites, is probably related to the risk of false splice site recognition. The low frequency of AC is due to a decreased abundance of both A and C.

The most frequent dinucleotides outside transitional zone [–5, 6] were TG, CT, and CC in both introns and exons; GG and TT in introns; and CA, AG, and AA in exons. Within transitional zone [–5, 6] (Fig. 5b), frequency peaks were observed for AG (positions –2 and 3), AA (positions –3 and 2), CA (position –3), GG (position –1), and GT (position 4).

Table 3. Dinucleotide frequencies in the [–40, –2] and [2, 40] donor site neighborhoods

Dinucleotide	Exon	Intron
AA	124	99
AC	89	65
AG	124	103
AT	85	91
CA	135	94
CC	117	113
CG	48	36
CT	113	116
GA	125	91
GC	106	97
GG	108	148
GT	67	99
TA	54	80
TC	95	91
TG	120	127
TT	89	151

Conditional dinucleotide frequencies (computed according to the above equation) in the donor site neighborhood are characterized in Table 4. The complete data are available at http://www.imb.ac.ru/splicingsites/data.zip.

In particular, as evident from Table 4 and Fig. 6, the following dinucleotides were characterized by the highest deviations of the conditional frequency from the expected unity: CG, GT, TA (low), CC, and GG (high) in the intron and CG, TA (low), CT, and TG (high) in the exon.

Similarly to the acceptor site, the variance in the intron was higher than in the exon (0.13 vs. 0.10). The conditional frequency charts for all dinucleotides are shown in Fig. 6.

In zone [–5, 6], CG stands out with an extremely low abundance, while the CA, CC, and CT levels are considerably increased as a consequence; the situation with TA can be interpreted in the same way.

Comparison to the complete genome. Comparing our data on the dinucleotide composition of the splice site neighborhoods and similar data earlier obtained for human genomic sequences [22], we detected some features characteristic of the splice site neighborhood and differing from the average human genome statistics. The conditional dinucleotide frequencies nearby splice sites were normalized to genomic conditional frequencies (Fig. 7).

The other results are available at http://www.imb.ac.ru/splicingsites/data.zip.

The analysis of the frequencies revealed the following patterns.

Firstly, there is a substantial difference in CG conditional frequency, which reaches 0.40 nearby splice sites as opposed to genomic 0.23. When only exon neighborhoods are considered, the difference is more than twofold.

Secondly, the acceptor dinucleotide AG is in deficit in the intron acceptor site neighborhood compared to the total genome (the conditional frequencies are 1.08 and 1.18, respectively).

Thirdly, AA is considerably more abundant in the intron and less abundant in the exon neighborhoods of splice sites.

Fourthly, TA and GG are in local deficit in exon neighborhoods (by the frequency compared to the genomic average).

Next, the conditional frequencies of GA and TG are increased in the splice site neighborhood.

Finally, the frequencies of AA, CC, and GG in the intron splice site regions are higher than the genomic average and lower in the exon regions.

DISCUSSION

Principal explicit (i.e., not hidden in various model parameters) data on the occurrence of different nucleotides and dinucleotides had already been obtained in the 1980s (see [21] for a review). It has been shown that, in the intron neighborhood of acceptor sites, T occurs more frequently and A and G are less frequent than expected for random distribution. Shapiro–Senapathy PSWMs have been obtained for nucleotides and dinucleotides [4], including a description of the AG and GT shadows in exons. Note that, under consideration of these shadows, both splice sites contain the same conserved motif GTAG. Its presence can be explained as either evolutionary (it may have been a splice protosite) or by the features of nucleotide recognition during splicing. The occurrence of the conserved dinucleotides has recently been investigated [14–16].

We systematically studied the positional frequencies of all nucleotides and dinucleotides in the neighborhood of splice sites. The study was carried out with a sample of over 190 000 human splice sites contained in the EDAS database [17]; the authenticity of each site was verified by comparing genomic sequences with mRNAs and ESTs during the database compilation. Owing to the volume of the database analyzed, the results can be considered highly reliable and the quantitative differences valid, because the statistical error is estimated at less than $\sim 1/\sqrt{N}$, where N is the number of DNA sequences analyzed. In the present work, N = 192 557 and the error is less than 0.005. Our

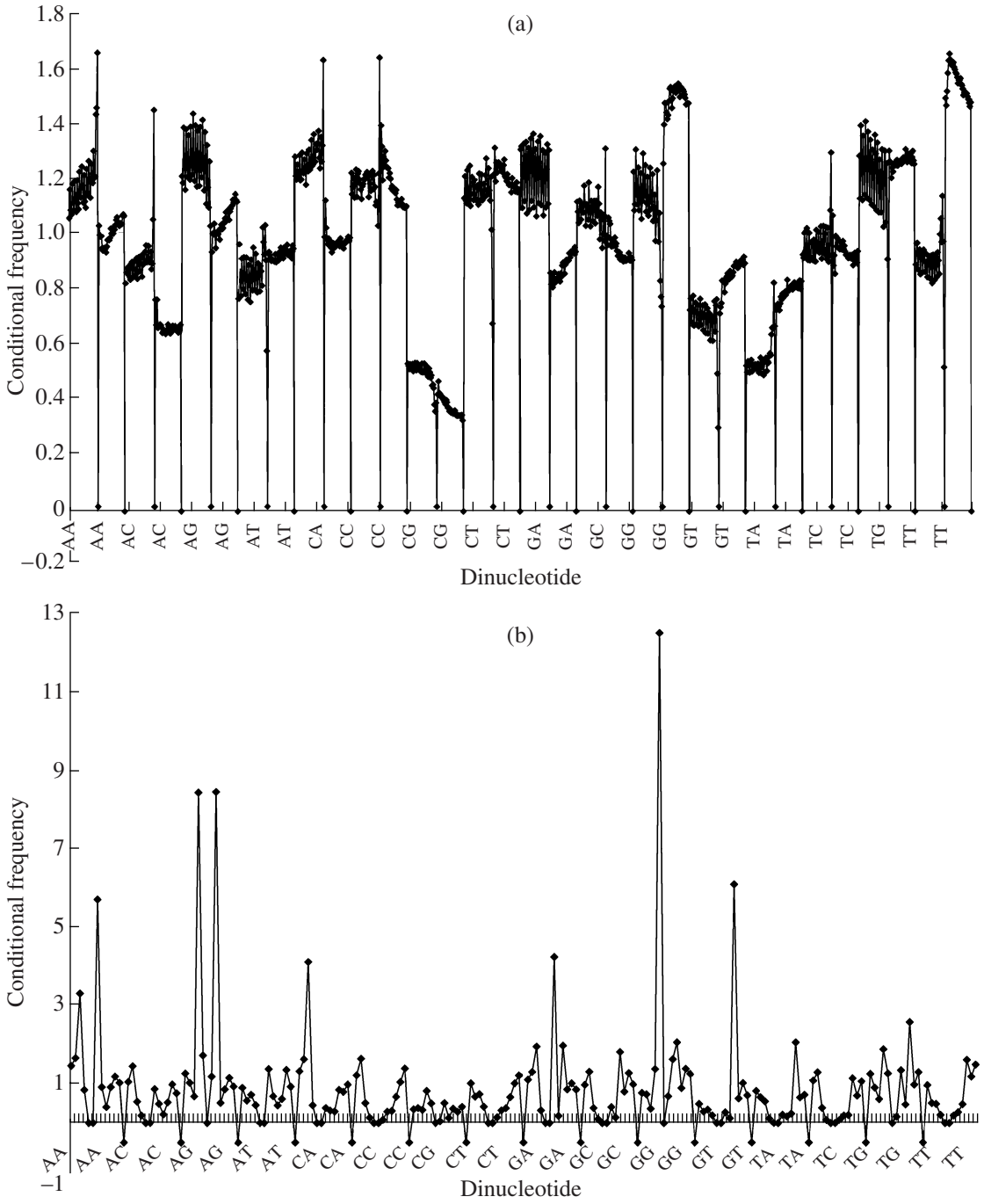

Fig. 5. Dinucleotide frequencies in the donor site neighborhood in the sample of 192 557 fragments: (a) distant neighborhood (positions [–40, -4] and [8, 39]) and (b) nearest neighborhood (positions [–3, –1] and [2, 7]). Each figure includes 2×16 charts separated with vertical lines. There are two charts for each dinucleotide: the left one represents exon positions (with negative coordinates) and the right one is for intron positions (with positive coordinates). The lines separating the exon and intron charts for the same nucleotide are marked on the X axis. The lines separating the charts for different nucleotides are marked below the X axis. Y axis, the ratio of the total number of a dinucleotide at a given position to the random average of 192 557/16.

results agree with the previous data, providing additional and more specific information. In particular, the following facts are worth noting.

(1) In the intron neighborhood of the acceptor site, the region of a T excess is nearly the same as the region of G deficiency; the same holds for the regions of a C excess and A deficiency.

(2) The region of a T excess (and G deficiency) in the intron neighborhood of the acceptor site is considerably larger than the region of a C excess (and A deficiency).

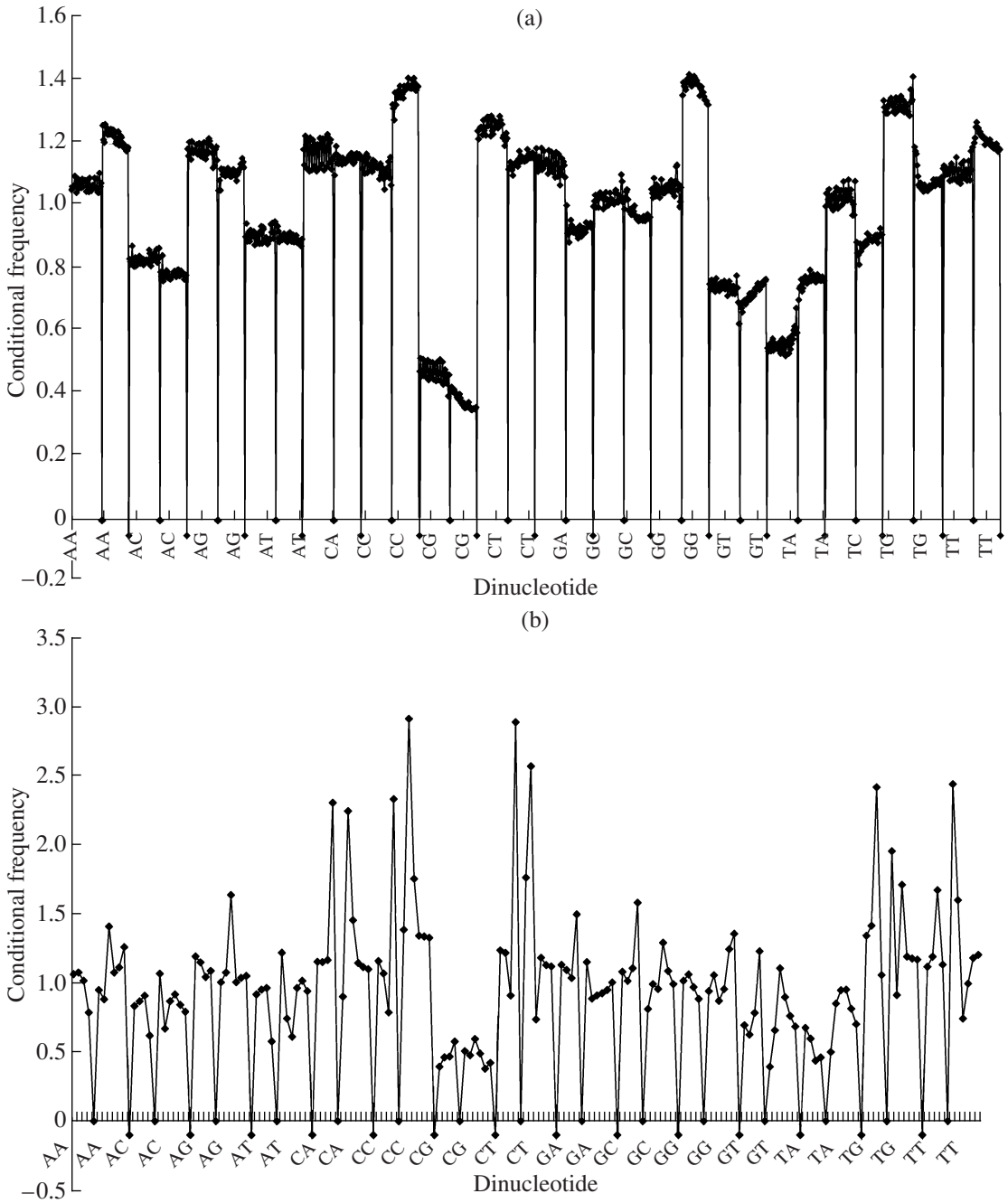

Fig. 6. Conditional dinucleotide frequencies (related to the product of the component nucleotide frequencies) nearby donor sites in the sample of 192 557 fragments: (a) distant neighborhood (positions [–40, –4] and [7, 39]) and (b) nearest neighborhood (positions [–3, –1] and [2, 6]). Data are presented as in Fig. 5.

(3) The 3-nt periodicity of nucleotide and dinucleotide frequencies was more accurately described. In particular, the amplitude of the C frequency variation in the exon neighborhoods of the acceptor and donor sites is considerably lower than that of the other nucleotides.

(4) Dinucleotide frequency variations cannot be reduced to variations of the component nucleotide frequencies: the relative dinucleotide frequency varies among individual positions from 0.3 to 2.4 in the acceptor site neighborhood and from 0.35 to 2.9 in the donor site neighborhood.

(5) The specific features of the dinucleotide distribution characterized in different regions of the splice site neighborhoods include an AG deficit in the nearest intron acceptor site region; a TG preference in the

acceptor site neighborhood; a preference of GG and TT in the intron region and of CA, AG, and AA in the exon region of donor sites; a GT and TA deficit and a TG and CT excess in the donor site neighborhood; and a CG and TA deficit in the splice site neighborhood.

(6) Comparative analysis of the conditional dinucleotide frequencies nearby splice sites and over the human genome identified CG, AG, AA, and TA as the dinucleotides characteristically preferred in the splice site neighborhood.

As our work was aimed at detecting the characteristic features of nucleotide and dinucleotide compositions of the flanking sequences, but not at explaining them biophysically, we shall not now discuss the problem of CG methylation and other hypotheses concerning the causes of dinucleotide preference or deficiency [22, 23].

Comparative analysis of the dinucleotide statistics in the human genome and nearby splice sites provides a basis for further research. It suggests itself as the next step in studying the nucleotide composition of the splice site-flanking sequences that the splice site neighborhoods should be compared separately to coding sequences and intron regions. It seems advisable to perform such an analysis with genomic sequences from different organisms to determine species-specific patterns.

We hope that the above results will promote a better understanding of the splicing mechanism, although more experimental data will be required for their interpretation. The molecular mechanics of splicing is at present insufficiently studied to unambiguously associate any characteristics of genetic sequences nearby splice sites with a mechanism of their recognition by the spliceosome. But the small size of the sites themselves is, technically, unfavorable for the efficient recognition, characteristic of the spliceosome. It seems obvious that specific features of nucleotide sequence fragments including the nearest splice site neighborhood play a crucial role in splice site recognition. Therefore, every statistically distinctive feature that differentiates splice sites from the genomic context is a contribution to developing an efficient method of identifying the exon–intron gene structure. Under consideration of the molecular mechanism of the interaction between spliceosome proteins and nucleic acids, as well as the small dimensions characteristic of the interacting objects (a few nucleotides/amino acids), it seems reasonable that the features of short oligonucleotides (and even individual nucleotides) in the neighborhood of splice sites promote their discrimination from the total sequence. As shown in this work, nucleotide and dinucleotide frequencies nearby both donor and acceptor sites have well-defined characteristic features, which may be used in developing

Table 4. Minimum, maximum, mean, and variance for the conditional frequencies of individual dinucleotides in the intron and exon donor site neighborhoods

Intron				
dinucleotide	minimum	maximum	mean	variance
AA	0.88	1.41	1.20	0.09
AC	0.67	1.07	0.80	0.06
AG	1.00	1.64	1.11	0.09
AT	0.61	1.22	0.90	0.08
CA	0.90	2.24	1.18	0.19
CC	1.28	2.91	1.42	0.26
CG	0.35	0.59	0.39	0.05
CT	0.74	2.57	1.19	0.26
GA	0.89	1.15	0.93	0.04
GC	0.81	1.29	0.99	0.07
GG	0.87	1.42	1.33	0.14
GT	0.39	1.11	0.73	0.09
TA	0.50	0.95	0.77	0.07
TC	0.79	1.39	0.91	0.11
TG	0.91	1.95	1.12	0.18
TT	0.74	2.44	1.24	0.23
Exon				
dinucleotide	minimum	maximum	mean	variance
AA	0.79	1.11	1.06	0.05
AC	0.62	0.91	0.83	0.04
AG	1.04	1.22	1.17	0.03
AT	0.58	0.96	0.90	0.06
CA	1.11	2.30	1.20	0.18
CC	0.79	2.33	1.15	0.20
CG	0.39	0.58	0.47	0.03
CT	0.91	2.89	1.28	0.27
GA	1.04	1.50	1.14	0.07
GC	0.99	1.58	1.04	0.09
GG	0.89	1.13	1.05	0.04
GT	0.63	1.23	0.75	0.08
TA	0.44	0.68	0.56	0.04
TC	0.97	1.67	1.05	0.11
TG	1.06	2.42	1.35	0.18
TT	1.07	1.67	1.13	0.09

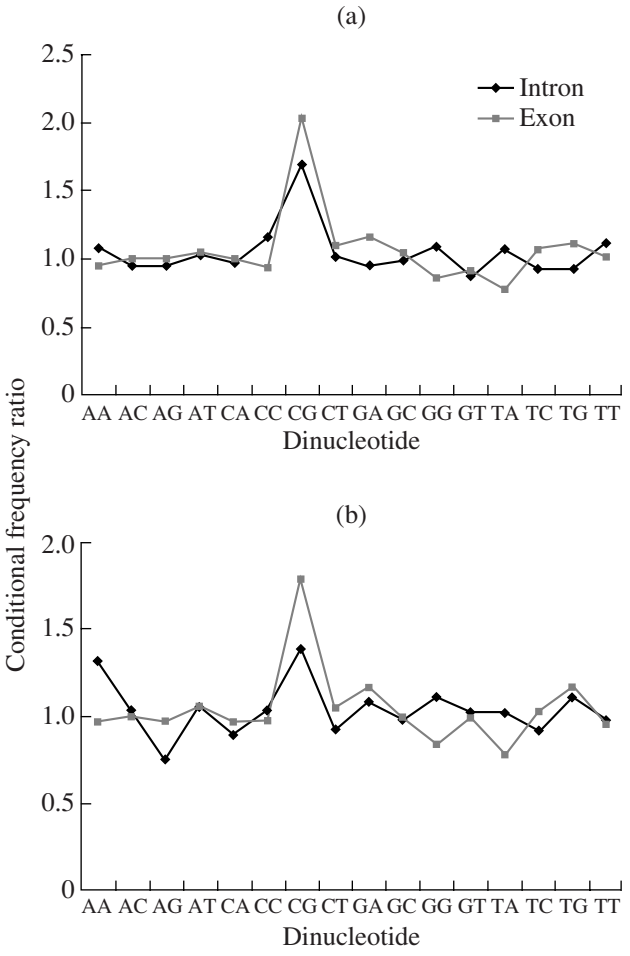

Fig. 7. Conditional dinucleotide frequencies normalized to genomic conditional frequencies in the splice site neighborhood. Y axis, the ratio of conditional dinucleotide frequencies in the splice site neighborhood to conditional dinucleotide frequencies calculated over the genome. (a) Donor site and (b) acceptor site.

new methods of splice site recognition and gene structure annotation.

ACKNOWLEDGMENTS

We are grateful to R. Nurtdinov and A.A. Mironov for their help in compiling the sample set of splice sites and to M.S. Gelfand, A.A. Mironov, and M.Yu. Borodovskij for useful discussion of the results.

The work was supported by the Russian Foundation for Basic Research (project no. 06-04-49249), INTAS (grant no. 05-1000008-8028), the program Molecular and Cell Biology of the Presidium of the Russian Academy of Science (project no. 10), and FIRCA NIH (project no. R03 TW005899).

REFERENCES

1. Michelle L., Hastings M.L., Krainer A.R. 2001. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol*. **13**, 302–309.
2. Burset M., Seledtsov A., Solovyev V.V. 2000. Analysis of canonical and non-canonical splite sites in mammalian genomes. *Nucleic Acids Res*. **28**, 4364–4375.
3. Sheth N., Roca X., Hastings M. L., Roeder T., Krainer A.R., Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*. **34**, 3955–3967.
4. Shapiro M.B., Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Res*. **15**, 7155–7174.
5. Hebsgaard S.M., Korning P.G., Tolstrup N., Engelbrecht J., Rouze P., Brunak S. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res*. **24**, 3439–3452.
6. Collins L., Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol*. **22**, 1053–1066.
7. Carmel I., Tal S., Vig I., Gil A. 2004. Comparative analysis detects dependencies among the 5' splice-site positions. *RNA*. **10**, 828–840.
8. Vorechovsky I. 2006. Aberrant 3' splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res*. **34**, 4630–4641.
9. Claverie J.M., Audic S. 1996. The statistical significance of nucleotide position-weight matrix matches. *CABIOS*. **12**, 431–439.
10. Lukashin A.V., Borodovsky M. 1999. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res*. **26**, 1107–1115.
11. Xu Y., Einstein J.R., Mural R.J., Shah M., Uberbacher E.C. 1994. An improved system for exon recognition and gene modeling in human DNA sequences, published presentation. *The Second International Conference on Intelligent Systems for Molecular Biology.* Stanford University, San Francisco, CA, USA.
12. Churbanov A., Rogozin I.B., Jitender S.D., Hesham A. 2006. Method of predicting splice sites based on signal interactions. *Biol. Direct*. **1**, 10.
13. Galperin M.Y. 2006. The molecular biology database collection: 2007 update. *Nucleic Acids Res*. **35**, D3–D4.
14. Eskesena S.T., Eskesen F.N., Ruvinskya A. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics*. **167**, 543–550.
15. Kraloviova J., Christensen M.B., Voechovsk I. 2005. Biased exon/intron distribution of cryptic and de novo 3' splice sites. *Nucleic Acids Res*. **33**, 4882–4898.
16. Roca X., Sachidanandam R., Krainer A.R. 2003. Intrinsic differences between authentic and cryptic 5' splice sites. *Nucleic Acids Res*. **31**, 6321–6333.

17. Neverov A.D., Artamonova I.I., Nurtdinov R.N., Frishman D., Gelfand M.S., Mironov A.A. 2005. Alternative splicing and protein function. *BMC Bioinform*. **6**, 266.

18. Gelfand M.S. 1985. Statistical analysis of mammalian pre-mRNA splicing sites. *Nucleic Acids Res*. **10**, 6369–6382.

19. Duret L., Galtier N. 2000. The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol*. **17**, 1620–1625.

20. Boudraa M., Perrin P. 1987. CpG and TpA frequencies in the plant system. *Nucleic Acids Res*. **15**, 5729–5737.

21. Zhang M.Q. 1998. Statistical features of human exons and their flanking regions.1998. *Human Mol. Genet*. **7**, 919–932.

22. Gentles A.J., Karlin S. 2001. Genome-scale compositional comparisons in eukaryotes. *Genome Res*. **11**, 540–546.

23. Grantham R. 1980. Workings of the genetic code. *Trends Biochem. Sci*. **5**, 327–331.