ISSN 0006-3509, Biophysics, 2006, Vol. 51, No. 4, pp. 515–522. © Pleiades Publishing, Inc., 2006.
Original Russian Text © E.O. Ermakova, D.B. Mal'ko, M.S. Gelfand, 2006, published in Biofizika, 2006, Vol. 51, No. 4, pp. 581–588.

MOLECULAR BIOPHYSICS

# Evolutionary Differences between Alternative and Constitutive Protein-Coding Regions of Alternatively Spliced Genes of *Drosophila*

**E. O. Ermakova[a,b], D. B. Mal'ko[c], and M. S. Gelfand[a,b,c]**

[a]*Lomonosov Moscow State University, Moscow, 199992 Russia*
[b]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia*
[c]*State Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 117545 Russia*

Received October 20, 2005; in final form, November 7, 2005

**Abstract**—A total of 790 *Drosophila melanogaster* genes that are alternatively spliced in a coding region and have orthologs in *Drosophila pseudoobscura* were studied. It proved that nucleotide substitutions are accumulated in alternative coding regions more rapidly than in constitutive coding regions. Moreover, the evolutionary patterns of alternative regions differing in insertion–deletion mechanisms (use of alternative promoters, splicing sites, or polyadenylation sites) differ significantly. The synonymous substitution rate in coding regions of genes varies more strongly than the nonsynonymous substitution rate. The patterns of substitutions in different classes of alternative regions of *Drosophila melanogaster* and mammals differ considerably.

DOI: 10.1134/S0006350906040014

*Key words:* molecular evolution, alternative splicing, *Drosophila*, nucleotide substitutions

## INTRODUCTION

Alternative splicing is one of the main mechanisms creating the proteomic diversity in eukaryotes. Only 80% of alternative exons of the fruit fly *Drosophila melanogaster* are conserved in *Drosophila pseudoobscura*, and 50% thereof are conserved in the mosquito *Anopheles gambiae*, whereas the percentages of conserved constitutive exons are 97 and 77%, respectively [1]. One alternatively spliced gene codes for several proteins with common structural segments. These proteins may perform different functions or the same function under different conditions (e.g., they may be expressed in certain tissues and/or certain developmental stages of the organism).

Since it is alternative coding regions of pre-mRNA, i.e., regions that can be eliminated or conserved in processing, and also coding regions located between alternative promoters or polyadenylation sites that are responsible for the specificity of protein isoforms, it is natural to expect that the corresponding regions of chromosomes undergo a different evolution than chromosomal regions corresponding to constitutive coding regions, i.e., regions that are always present in a processed mRNA. A number of studies corroborate this hypothesis. Alternative splicing sites are weaker than constitutive ones [2]. Noncanonical GC–AG introns are more often alternative than canonical GT–AG introns [3]. Among exons common to human and mouse, for 77% of alternative cassette exons and only for 17% of constitutive exons the neighboring long sequences of introns are also conserved [4].

The evolution of a gene proceeds by fixation of mutations (nucleotide substitutions) and its participation in chromosomal rearrangements, including the loss and gain of exons and introns. The evolution of the exon–intron structure of alternatively spliced genes of *D. melanogaster* was studied earlier [1]. In

particular, it was shown that new introns more frequently arise in alternative regions of a gene than in constitutive regions.

It is natural to assume that nucleotide substitutions leading to amino acid substitutions in protein are more rapidly accumulated in alternative regions. Iida and Akashi [5] analyzed 26 pairs of alternatively spliced human genes and their orthologs in other mammals and showed that the nonsynonymous substitution rate in alternative regions of genes is higher than in constitutive regions, and vice versa, the synonymous substitution rate in alternative regions is lower than in constitutive regions. However, our analysis of 3029 orthologous pairs of alternatively spliced genes of human and mouse demonstrated a higher rate of substitutions of both types in alternative coding regions as compared with constitutive regions [6]. In this work, we analyze nucleotide substitutions in 790 alternatively spliced genes of *D. melanogaster*, comparing them with orthologs in *D. pseudoobscura*.

Three main types of alternative splicing are distinguished [7]. In type I splicing, different promoters of the same gene are used to obtain cognate pre-mRNAs with different 5′-proximal regions of different lengths. As a result, proteins with different N-termini can form. Alternative type II splicing involves pre-mRNAs with different 3′-proximal sequences of different lengths, which generally results from polyadenylation of transcripts at different sites. Correspondingly, proteins with different C-termini can be produced. And finally, in alternative type III splicing, identical pre-mRNA are processed into different functional mRNAs. In splicing, some or other alternative splicing sites are chosen.

The existence of several types of alternative splicing in eukaryotes stimulated us to separately study the evolution of DNA regions corresponding to N-terminal alternative (N-alternative), C-terminal alternative (C-alternative), and inner alternative (I-alternative) regions of protein. It proved that the evolutionary behaviors of the three classes of alternative regions in Drosophila differ significantly (see Results and Discussion).

## PROCEDURE

Genomic DNA and protein isoforms of *D. melanogaster* were taken from the FlyBase database, version 3 (ftp://flybase.net/genomes/Drosophila_melanogaster/), and genomic DNA of *D. pseudoobscura* was obtained from The Baylor College of Medicine Human Genome Sequencing Center (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Dpseudoobscura). The orthology between genes of *D. melanogaster* and *D. pseudoobscura* was established as described before [1]. The protein isoforms of *D. melanogaster* were mapped on the genome using the program Pro-Frame [8]. We identified a total of 790 *D. melanogaster* genes meeting the following requirements: the existence of different protein isoforms, the existence of a unique ortholog in *D. pseudoobscura*, and the existence of alternative regions conserved in *D. pseudoobscura* (the conservation of a region meant the existence of alignment that was significant according to published criteria [1] and the conservation of splicing sites).

### Analysis Algorithm and Meta-Alignments

Data analysis consisted of the following steps. First, we separated subsamples of genes (in particular, genes with long alternatives were considered separately, see below) and classes of constitutive and alternative regions of alignments of these genes. Second, meta-alignments (concatenated alignments) of regions of separated classes were composed. We considered meta-alignments of two types: local meta-alignments combining coding regions of the same class within a single gene and global meta-alignments combining regions of the same class of all genes of a certain subsample (Table 1). Incomplete codons and codons with deletions in alignment were not included in meta-alignments. Third, we evaluated the similarity of the protein products and the synonymous and nonsynonymous nucleotide substitution rates in genome regions of a given class (Table 2; Fig. 1). This evaluation required the existence of sufficiently long alignment [9]; we used a threshold of 80 nt. The use of meta-alignments allowed us to take into account not only long alternative segments of genes, such as cassette exons, but also very short ones, such as extensions, i.e., regions between two alternative donor sites or two acceptor sites.

### Considered Evolutionary Parameters

For each meta-alignment, we estimated four evolutionary parameters. The fraction *Id* of coinciding amino acids in translated meta-alignment enables one

**Table 1.** Structure of global meta-alignments

| Name of sample of genes | Description of subsample of genes | Number of genes in subsample | Description of subsample of coding regions | Meta-alignment length, nt |
|---|---|---|---|---|
| Reference sample | All alternatively spliced genes | 790 | N-alternative | 266935 |
| | | | I-alternative | 84558 |
| | | | C-alternative | 183581 |
| | | | All alternative | 535074 |
| | | | Constitutive | 1334760 |
| Genes with long alternatives | Genes in which total alignment length of alternative regions exceeds 80 nt and total alignment length of constitutive regions exceeds 80 nt | 588 | Alternative | 477229 |
| | | | Constitutive | 940465 |
| Genes with long N-alternatives | Genes in which total alignment length of N-terminal alternative regions exceeds 80 nt and total alignment length of constitutive regions exceeds 80 nt | 308 | N-alternative | 247294 |
| | | | Constitutive | 451872 |
| Genes with long I-alternatives | Genes in which total alignment length of inner alternative regions exceeds 80 nt and total alignment length of constitutive regions exceeds 80 nt | 145 | I-alternative | 75271 |
| | | | Constitutive | 342932 |
| Genes with long C-alternatives | Genes in which total alignment length of C-terminal alternative regions exceeds 80 nt and total alignment length of constitutive regions exceeds 80 nt | 184 | C-alternative | 150755 |
| | | | Constitutive | 240306 |

to estimate the current divergence of regions of homologous proteins. The nonsynonymous substitution rate $d_N$ also serves as a measure of the divergence of amino acid sequences corresponding to homologous regions of two genes, but, along with this, $d_N$ characterizes the "saturation" of a coding region with nonsynonymous substitutions. The synonymous substitution rate $d_S$ allows one to judge about both the intensity of mutations in one or another coding region (in comparison with $d_N$) and the evolution of "nonprotein" elements of the gene, e.g., regulatory sequences. The normalization of $d_N$ and $d_S$ is matched, and whereas $d_N$ and $d_S$ estimate the number of nucleotide substitutions counted from the moment of divergence of two organisms, their ratio $\omega = d_N/d_S$ is already not a function of time but a characteristic of the selection pressure on this region. In particular, at $\omega > 1$, it is concluded that the coding region is under positive selection.

### Estimation of Nucleotide Substitution Rate. The Ina Method

We estimated $d_N$ and $d_S$ by the Ina method [9]. This method belongs to "evolutionary pathway" methods, in which, for each pair of alignment codons, all the shortest pathways of transformation of one codon into the other by single-nucleotide substitutions are considered.

The genetic code is degenerate. A substitution for a codon as a result of a mutation of a nucleotide can be synonymous, if the initial triplet and the triplet obtained by the mutation code for the same amino acid. Each position of a nonterminating codon has synonymous potential $s$ and nonsynonymous potential $n$, such as $s + n = 1$. In the general case, (non)synonymous potential of a position in a codon is the probability to obtain a (non)synonymous substitution by a mutation of the nucleotide in this position. If a substitution for the base in one of the positions of a codon (the other positions being fixed) leads to a

**Table 2.** Comparison of the evolutionary parameters of alternative and constitutive regions in terms of global meta-alignments

| Sample of genes | Class of coding regions | *Id* | $d_N$ | $d_S$ | ω |
|---|---|---|---|---|---|
| Reference sample | N-alternative | 0.66 | 0.23 | 0.36 | 0.62 |
| | I-alternative | 0.54 | 0.33 | 0.23 | 1.43 |
| | C-alternative | 0.61 | 0.25 | 0.28 | 0.89 |
| | All alternative | 0.62 | 0.25 | 0.31 | 0.80 |
| | Constitutive | 0.65 | 0.22 | 0.27 | 0.81 |
| Genes with long alternatives | Alternative | 0.62 | 0.25 | 0.31 | 0.81 |
| | Constitutive | 0.65 | 0.22 | 0.26 | 0.84 |
| Genes with long N-alternatives | N-alternative | 0.66 | 0.22 | 0.36 | 0.62 |
| | Constitutive | 0.63 | 0.23 | 0.26 | 0.90 |
| Genes with long I-alternatives | I-alternative | 0.53 | 0.33 | 0.22 | 1.47 |
| | Constitutive | 0.63 | 0.23 | 0.21 | 1.06 |
| Genes with long C-alternatives | C-alternative | 0.62 | 0.25 | 0.27 | 0.93 |
| | Constitutive | 0.68 | 0.20 | 0.32 | 0.63 |

*Id* is the fraction of coinciding amino acids in translated meta-alignment;
$d_N$ is the rate of nonsynonymous substitutions of nonsynonymous positions;
$d_S$ is the rate of synonymous substitutions of synonymous positions.
$\omega = d_N/d_S$

nonsynonymous substitution for the codon, this position is called nonsynonymous and, for it, $s = 0$ and $n = 1$. And if any substitution for the base in this positions leads to a synonymous substitution for the codon, this position is called synonymous and, for it, $s = 1$ and $n = 0$. In the general case, $s$ and $n$ depend on the probabilities $\lambda_{XY}$ of mutation of nucleotide X into nucleotide Y in a unit time (X, Y = A, C, T, G). In the Ina method, the Kimura two-parameter model is accepted [10]: $\lambda_{TC} = \lambda_{CT} = \lambda_{AG} = \lambda_{GA} = \alpha$ and $\lambda_{AC} = \lambda_{CA} = \lambda_{CG} = \lambda_{GC} = \lambda_{AT} = \lambda_{TA} = \lambda_{GT} = \lambda_{TG} = \beta$ (Fig. 2). In this case, for each position of each codon $s$ and $n$ are readily expressed through $R = \alpha/\beta$ (Fig. 3). Counting nucleotide substitutions in coding regions of genes, the number of nonsynonymous substitutions is usually normalized to the total nonsynonymous potential of this region and the number of synonymous substitutions is normalized to the total synonymous potential of this region.

Over all the pairs of codons, the number of nucleotide differences of four types is summed: synonymous and nonsynonymous transitions and transversions (a transition is a substitution of a pyrimidine base for a pyrimidine base, or a substitution of a purine base for a purine base, and its created difference; and a transversion is a substitution of a purine base for a pyrimidine base, or a substitution of a pyrimidine base for a purine base, and its created difference; Fig. 2). If two codons differ in two or three positions, then, for each of the four types of substitutions, the arithmetic mean of their number over all possible minimal pathways of transformation of one codon into the other is taken. The minimal pathways containing termination codons are ignored. In particular, if none of the minimal pathways contains termination codons, then, for differences in two positions, we have two pathways, and for differences in three positions, we have six pathways (Fig. 2).

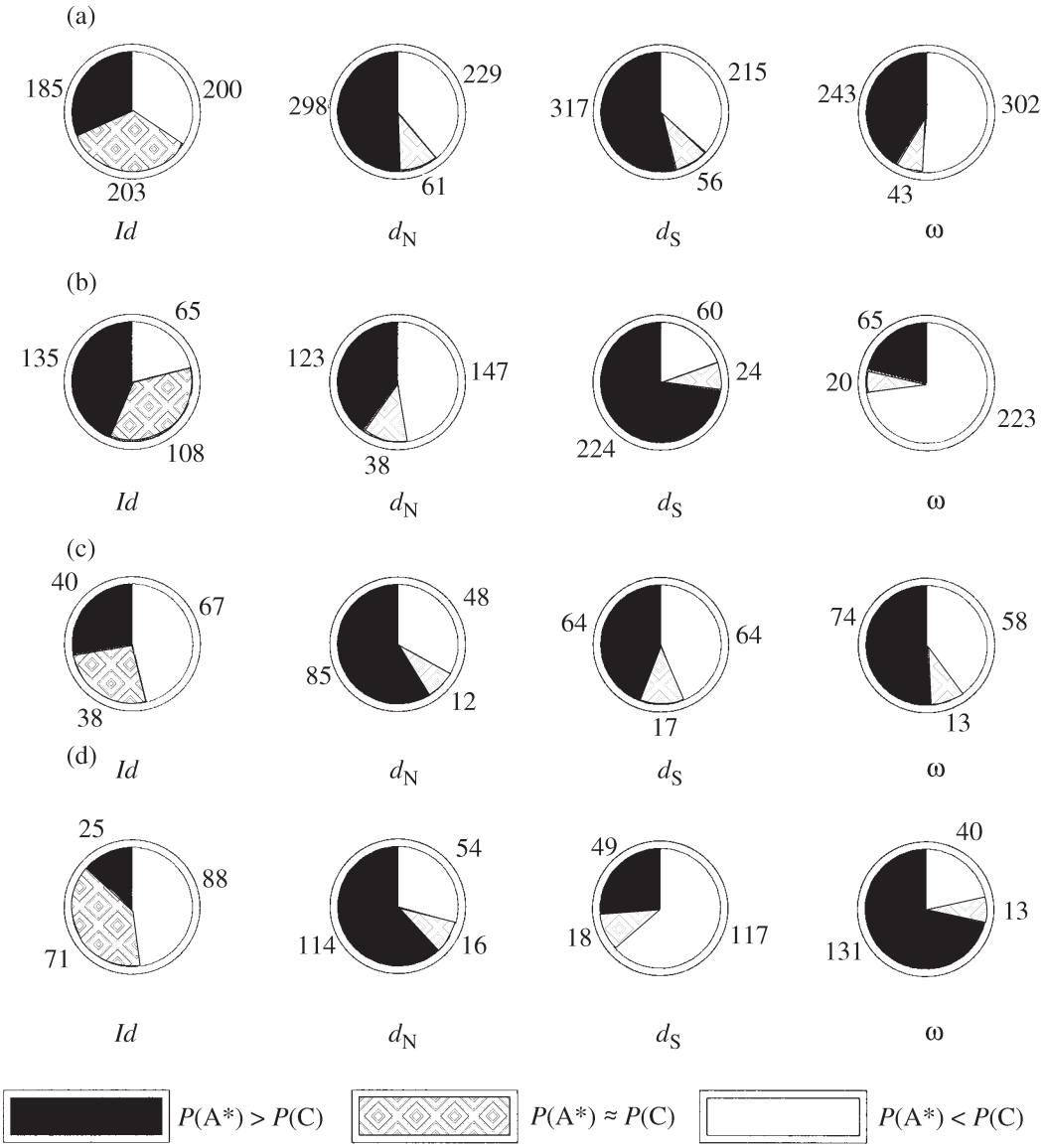


**Fig. 1.** Comparison of the evolutionary parameters of alternative and constitutive regions in terms of global meta-alignments. Each diagram describes a $P$(A*) vs. $P$(C) experiment, where $P$ is one of the evolutionary parameters (*Id*, $d_N$, $d_S$, or ω) and A* is a class of alternative regions (A, $A^N$, $A^I$, or $A^C$). A difference in a parameter $P$ between regions of classes A* and C was considered significant if $|P(A^*) - P(C)| / |P(A^*) + P(C)| > 0.05$. *For each sector, the number of genes in the corresponding group is indicated. (a) Genes with long alternatives. All alternative (A) vs. constitutive (C). (b) Genes with long N-alternatives. N-alternative ($A^N$) vs. constitutive (C). (c) Genes with long I-alternatives. I-alternative ($A^I$) vs. constitutive (C). (d) Genes with long C-alternatives. C-alternative ($A^C$) vs. constitutive (C).

Let $S^*$ be the arithmetic mean of the total synonymous potentials of aligned sequences, $S^*_{Ts}$ be the number of transitions observed in alignment, and $S^*_{Tv}$ be the number of transversions observed in alignment. Then, the observed frequencies of synonymous differences—transitions and transversions—in synonymous positions are equal to

$$P^*_S = \frac{S^*_{Ts}}{S^*} \qquad Q^*_S = \frac{S^*_{Tv}}{S^*}.$$

An estimate $d^*_S$ for $d_S$ is made by applying the Kimira correction for multiple substitutions to $P^*_S$ and $Q^*_S$ [10]:

$$d^*_S = -\frac{1}{2}\ln(1 - 2P^*_S - Q^*_S) - \frac{1}{4}\ln(1 - 2Q^*_S).$$

An estimate $d^*_N$ for $d_N$ is made similarly. The parameter ω is estimated as $d^*_N/d^*_S$.

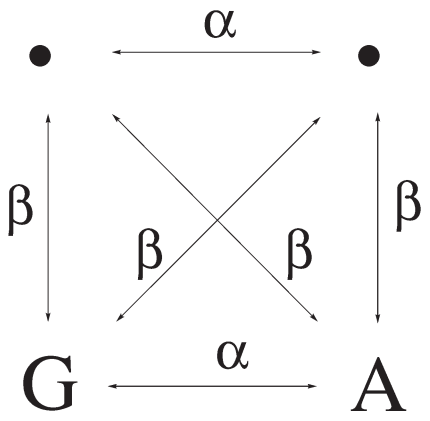


**Fig. 2.** Kimura two-parameter model of DNA evolution [10]: $\lambda_{TC} = \lambda_{CT} = \lambda_{AG} = \lambda_{GA} = \alpha$ and $\lambda_{AC} = \lambda_{CA} = \lambda_{CG} = \lambda_{GC} = \lambda_{AT} = \lambda_{TA} = \lambda_{GT} = \lambda_{TG} = \beta$.

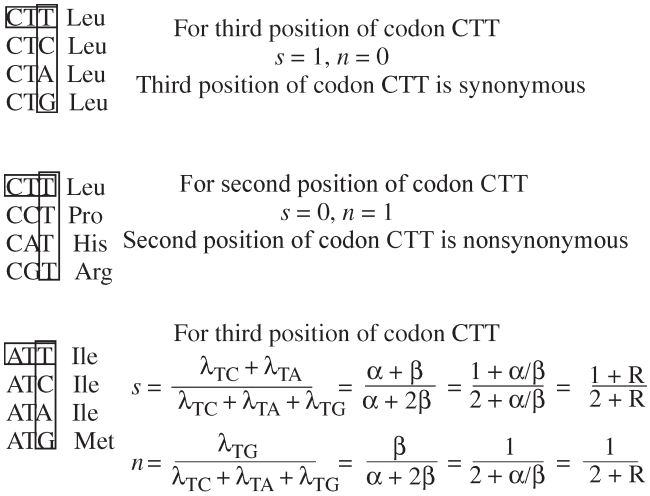


**Fig. 3.** Calculation of the synonymous potential *s* and the nonsynonymous potential *n* of a position in a codon in the Ina method [9].

### Asymmetry of Mutations of Nucleotides

Mutations of pyrimidine bases into pyrimidine bases or mutations of purine bases into purine bases in genomic DNA are much more frequent than mutations of pyrimidine bases into purine bases and vice versa. The ratio $R$ of the numbers of transitions and transversions, which is necessary to calculate the synonymous and nonsynonymous potentials of a nucleotide position, was estimated from meta-alignment of all regions of all of the 790 genes. If $P_3^*$ and $Q_3^*$ are the observed frequencies of transitions and transversions, respectively, in the third positions of alignment codons, then we suppose [9]

$$R = \frac{2\ln(-2P_3^* - Q_3^*)}{\ln(1-2Q_3^*)} - 1.$$

In our problem, $R = 2.24$.

### Analysis of Local Meta-Alignments

For genes with long alternatives, we compared the evolutionary behaviors of different alternative regions of different classes and constitutive regions using local meta-alignments (Fig. 1). A difference in a parameter $P$ between regions of classes X and Y was considered significant if $|P(X) - P(Y)|/|P(X) + P(Y)| > 0.05$.

## RESULTS AND DISCUSSION

We considered 790 alternatively spliced genes of *D. melanogaster* that have orthologs in *D. pseudoobscura*. The fraction *Id* of coinciding amino acids, the nonsynonymous substitution rate $d_N$, the synonymous substitution rate $d_S$, and $\omega = d_N/d_S$ in alternative (A) and constitutive (C) regions of genes were estimated (Table 2; Fig. 1). It proved that, on the average, $Id(A) < Id(C)$, $d_N(A) > d_N(C)$, $d_S(A) > d_S(C)$, and $\omega(A) < \omega(C)$. Thus, nucleotide substitutions are accumulated more rapidly in alternative regions. This suggests weakening of negative selection and/or strengthening of positive selection in alternative regions of genes. Our observations are consistent with the inverse dependence of the number of nucleotide substitutions in a coding region of a gene on the number of tissues in which the gene is expressed [11] and on its expression level [12] if it is assumed that constitutive regions are expressed in a larger number of tissues than alternative regions are and the expression level of the former is higher.

We also separately considered N-alternative ($A^N$), I-alternative ($A^I$), and C-alternative ($A^C$) regions of genes. For global meta-alignments of the base sample (Table 2), $Id(A^N) > Id(A^C) > Id(A^I)$ and $d_N(A^N) < d_N(A^C) < d_N(A^I)$; i.e., N-terminal alternative regions are most conserved and inner regions are least conserved. Moreover, $\omega(A^I) > 1$. This suggests that inner alternative regions of genes are under positive selection. In addition, $d_S(A^C) \approx d_S(C)$, $d_S(A^I) < d_S(C)$, and $d_S(A^N) > d_S(C)$. In coding regions of genes, several layers of information are coded, including information on protein sequences and regulatory sites. Nonsynonymous substitutions can change both of these layers, whereas synonymous substitution can

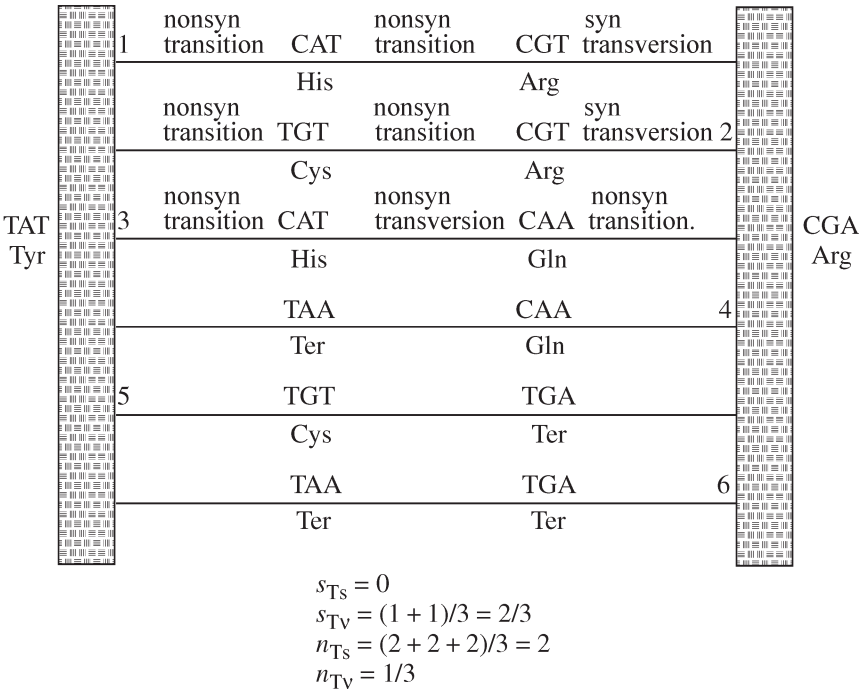


**Fig. 4.** Count of substitutions of different types between the codons TAT and CGA: $s_{Ts}$ is the number of synonymous transitions, $s_{Tv}$ is the number of synonymous transversions, $n_{Ts}$ is the number of nonsynonymous transitions, and $n_{Tv}$ is the number of nonsynonymous transversions. Between the codons TAT and CGA, there are 6 minimal evolutionary pathways, but three of them were ignored in calculations since they contain termination codons. Note that pathways 2 and 3 contain different numbers of synonymous and nonsynonymous substitutions.

change only the second of them. Our observations allow us to assume that, in N-alternative and I-alternative regions, two opposite scenarios are realized: in N-terminal alternatives, only changes in regulatory sites are favored ($d_N(A^N) \approx d_N(C)$, $d_S(A^N) > d_S(C)$), whereas in inner alternatives, only changes in amino acid sequences are ($d_N(A^I) > d_N(C)$, $d_S(A^I) < d_S(C)$). From this standpoint, C-terminal alternatives are intermediate. A similar analysis of meta-alignments of N-terminal, inner, and C-terminal constitutive regions revealed no differences in considered evolutionary parameters between these classes of coding regions of genes (unpublished data). Interestingly, in mammals, the distribution of substitutions in N-, I-, and C-alternative regions is quite different: the rates of substitutions of both types are much higher and the selection pressure on C-alternative regions is lower [6]. This is likely to be related to the expression level of alternative regions: it was shown for alternatively spliced genes of human that, the less frequently an alternative region is expressed, the higher the nonsynonymous substitution rate and the lower the synonymous substitution rate in it [13].

Since we considered only alternatives confirmed by proteins, some alternative regions may be included in a sample of constitutive regions; however, this could lead only to leveling of the observed differences between constitutive and alternative regions.

Among genes with long alternative regions of different classes, genes with long inner and C-terminal alternatives differ strongly from the general case. In such genes, the behavior of nonsynonymous substitutions is the same but the behaviors of synonymous nucleotide substitutions in coding regions of different classes differ significantly. In comparison with the reference sample, the synonymous substitution rate in constitutive regions of genes with long I-alternatives is much lower, whereas that in genes with long C-alternatives is considerably higher. Thus, the mechanism of insertion–deletion of alternative regions largely determines its evolution.

## ACKNOWLEDGMENTS

We thank A.S. Kondrashov, A.A. Mironov, R.N. Nurtdinov, and V.E. Ramenskii for fruitful discussions.

This work was supported in part by the Russian Foundation for Basic Research (project no. 04-04-49440), the Howard Hughes Medical Institute (grant no. 55000309), the Ludwig Institute for Cancer Research (CDRF RB0-1268), and also the Russian Science Support Foundation (M.S.G.) and the Program of the Russian Academy of Sciences "Molecular and Cell Biology".

## REFERENCES

1. D. B. Malko, in Proceedings of the Second International Moscow Conference on Computational Molecular Biology, 2005, p. 224.
2. F. Clark and T. A. Thanaraj, Hum. Mol. Genet. **11**, 451 (2001).
3. M. Burset, I. A. Seledtsov, and V. V. Solovyev, Nucl. Acids Res. **28**, 4364 (2000).
4. R. Sorek and G. Ast, Genome Res. **13**, 1631 (2003).
5. K. Iida and H. Akashi, Gene **261**, 93 (2000).
6. E. O. Ermakova, in Proceedings of the Second International Moscow Conference on Computational Molecular Biology, 2005, p. 95 .
7. M. Singer and P. Berg, *Genes and Genomes: A Changing Perspective* (University Science Books, Mill Valley, Calif., 1991; Mir, Moscow, 1998), Vol. 2.
8. A. A. Mironov, P. S. Novichkov, and M. S. Gelfand, Bioinformatics **17**, 13 (2001).
9. Y. Ina, J. Mol. Evol. **40**, 190 (1995).
10. M. Kimura, J. Mol. Evol. **16**, 111 (1980).
11. L. Duret and D. Mouchiroud, Mol. Biol. Evol. **17**, 68 (2000).
12. C. Pal, B. Papp, and L. D. Hurst, Genetics **158**, 927 (2001).
13. Y. Xing and C. Lee, Proc. Natl. Acad. Sci. USA **102**, 13526 (2005).