From: N. Kolchanov, and R. Hofestadt, eds. 2005. Bioinformatics Of Genome Regulation And Structure II, Springer. Pages 21-29 (The references were all collected at the end in the book)

Content sensors based on Codon structure and DNA methylation for gene finding in vertebrate genomes

Xuhua Xia
University of Ottawa, 150 Louis Pasteur, Ottawa, Canada; xxia@uottawa.ca

Abstract: All vertebrate genomes are heavily methylated at CpG dinucleotide sites, and methylated CpG dinucleotides are prone to CpG→TpG mutations through spontaneous deamination. This leaves different footprints on coding and non-coding sequences. We capture these different fingerprints by five indices that can be used to discriminate between coding and non-coding (intron) sequences. We also show that a linear discriminant function derived from a training set of coding and intron sequences from human chromosome 22 can be successfully used in gene-finding of the zebrafish genome.

Keywords: content sensor; DNA methylation; gene finding; vertebrate genome; codon structure

1. Introduction

There are two major categories of gene-finding methods. The first is based on known genes in molecular databases, and uses homology search by FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997). The second is based on known gene structures, and represented by GENSCAN (Burge and Karlin, 1997). Existing software for gene-finding often combine both approaches, e.g., GenMark (Hayes and Borodovsky, 1998), GLIMMER (Salzberg *et al.*, 1998), Orpheus (Frishman *et al.*, 1998), Projector (Meyer and Durbin, 2004) and YACOP (Tech and Merkl, 2003).

For structure-based prediction, the hidden Markov model (HMM) is frequently used in combination with the Viterbi algorithm (Baldi and Brunak, 2001; Pevzner, 2000). Whether HMM is effective depends much on whether the hidden states (e.g., exon, intron, tRNA, rRNA, intergenic sequence, etc.) emit different symbols (i.e., nucleotide combinations). Take for illustration the

classical HMM example of the dishonest casino dealer who switches between a fair die and a loaded die (i.e., two hidden states). If the loaded die differ much from the fair die, e.g., if the probability of having 6 is nearly 1 for the loaded die, then a short stretch of 6s is sufficient to identify the point of switching. If the loaded die differs only slightly from the fair die, then the switching event will be difficult to identify unless the dealer throws the loaded die consecutively for a long time. Similarly, if exons and introns differ dramatically in nucleotide combinations, then we would be able to distinguish between short exon and intron sequences. If they differ little, then we may not be able to tell them apart even with long sequences.

In vertebrate genomes, exons and introns are often quite short. So it is important to find features that differ substantially among exons, introns, RNA genes, etc. For this reason, extensive studies have been carried out to characterize the differences among different sequence states, leading to a variety of signal sensors such as the relatively uniform splicing sites (Foissac and Schiex, 2005; Gelfand, 1989; Tenney *et al.*, 2004) and the much less uniform exon-exon junctions (Gelfand, 1992), and content sensors such as unusual frequency distributions of words (Borodovsky and McIninch, 1993; Gelfand *et al.*, 1992; Pevzner *et al.*, 1989) that can be potentially used in gene-finding.

This paper focuses on content sensors that can discriminate between exons and introns, based on the sequence pattern created by DNA methylation. DNA methylation is a ubiquitous biochemical process particularly pronounced in vertebrate genomes (Bestor and Coxon, 1993; Rideout *et al.*, 1990; Sved and Bird, 1990). A typical representative of the vertebrate methyltransferase is the mammalian DNMT1 with five domains of which the NlsD, ZnD and CatD domains bind specifically to unmethylated CpG, methylated CpG and hemimethylated CpG sites, respectively (Fatemi *et al.*, 2001). Methylation of C in the CpG dinucleotide greatly elevates the mutation rate of C to T through spontaneous deamination of the resultant m^5C (Brauch *et al.*, 2000; Tomatsu *et al.*, 2004), generating strong footprints in both prokaryotic and vertebrate genomes (Xia, 2003, 2004). Here we develop indices to capture the differential methylation-mediated substitution patterns and nucleotide triplet structures between exons and introns for gene detection.

1.1. Nucleotide and dinucleotide frequencies by triplet sites

Designate the nucleotide frequencies of a sequence as p_A, p_C, p_G and p_T, and the sequence length as L. Consider both coding and non-coding sequences as a linear sequence of consecutive triplets, and the nucleotide frequencies at the three sites of the triplets as p_{i1}, p_{i2} and p_{i3}, where i = 1, 2, 3, and 4 corresponding to nucleotide A, C, G, T, respectively.

For non-exon (e.g., intron) sequences, there is no codon structure. So we expect $p_{i1} \approx p_{i2} \approx p_{i3} \approx p_i$, where p_i is the average of p_{i1}, p_{i2}, and p_{i3}. For coding sequences, the methylation will creates heterogeneity in nucleotide frequencies among the three sites. Take NCG codon for example, where N stands for any of the four nucleotides. DNA methylation and spontaneous deamination tend

to change these NCG codons to NTG and NCA codons (Note that CpG→TpG mutations in one DNA strand lead to CpG→CpA mutations in the complementary strand), with the former change being nonsynonymous and the latter synonymous. Because nonsynonymous substitution is generally much rarer than the synonymous substitutions in a large number of protein-coding genes in many organisms studied (Xia, 1998; Xia *et al.*, 1996; Xia and Li, 1998), we should expect NCG→NCA mutations more often than NCG→NTG mutations. This tends to increase the frequency of A at the third codon position. Similarly, dicodons such as "NNC GNN" tend to mutate synonymously to "NNT GNN" with DNA methylation, increasing the frequency of T at the third codon position. Thus, in contrast to non-coding sequences where we expect $p_{i1} \approx p_{i2} \approx p_{i3} \approx p_i$, we should expect $p_{i1} \neq p_{i2} \neq p_{i3} \neq p_i$ in coding sequences. This suggests that the deviation of p_{ij} (where j = 1, 2 and 3 corresponding to the three triplet sites) from p_i can contribute to the discrimination between coding and non-coding sequences. A measure of this deviation that is independent of L is as follows:

$$\varphi_{Nuc} = \frac{\sum_{i=1}^{4} \sqrt{\frac{\sum_{j=1}^{3} \frac{(f_{ij} - f_i)^2}{f_i}}{N_i}}}{4} \qquad (1)$$

where f_{ij} stands for the number of nucleotide i at codon position j, f_i is the mean number of nucleotide i averaged over the three codon (triplet) positions, and N_i is the sum of nucleotide i in the sequence. We expect φ_{Nuc} to be greater for coding sequences than for non-coding sequences.

Following a similar line of reasoning, we expect the dinucleotide frequencies at triplet positions (1,2), (2,3) and (3,1) to be similar to each other in non-coding sequences but different in coding sequences. Designate the number of dinucleotides as $f_{ij.k}$, where ij = AA, AC, ..., TT, respectively, and k = 1, 2, 3 corresponding to the triplet positions (1,2), (2,3) and (3,1), respectively. The deviation of $f_{ij.k}$ from f_{ij}, which is the number of dinucleotide i averaged over the three triplet positions, should also contribute to the discrimination between coding and non-coding sequences. A measure of this deviation that is independent of L is:

$$\varphi_{DiNuc} = \frac{\sum_{i=1}^{4} \sum_{j=1}^{4} \sqrt{\frac{\sum_{k=1}^{3} \frac{(f_{ij.k} - f_{ij})^2}{f_{ij}}}{N_{ij}}}}{16} \qquad (2)$$

For short sequences, N_{ij} may be zero, in which case φ_{DiNuc} is not defined, or very small, in which case φ_{DiNuc} would fluctuate widely. To avoid this problem, the computation can be done by setting valid $N_{ij} \geq 6$ and the denominator will be the number of valid N_{ij} values.

1.2. Differential methylation intensity

DNA methylation and deamination decrease the CG-containing triplets and increase the UG- and CA-containing triplets. However, their effect is stronger on introns than on coding sequences because of weaker selection constrains on introns than on coding sequences, e.g., all CGN→TGN, CGN→CAN and NCG→NTG mutations are nonsynonymous and should be selected against in coding sequences but not in non-coding sequences. For this reason, the intensity of methylation (designated I_m) should be greater in introns than in coding sequences:

$$I_m = \frac{(f_{NUG,UGN,NCA,CAN} - f'_{NUG,UGN,NCA,CAN}) - (f_{NCG,CGN} - f'_{NCG,CGN})}{f_{NUG,UGN,NCA,CAN} + f_{NCG,CGN}} \quad (3)$$

where f is the sum of frequencies of those subscripted codons, and f' is the corresponding expectation computed simply by

$$f'_{ijk} = N_{triplet} P_i P_j P_k \quad (4)$$

where $N_{triplet}$ is the total number of non-overlapping triplets in the sequence. A more reasonable expectation would be (by taking AAA and AAG for illustration):

$$f'_{AAA} = N_{Lys} \bullet \frac{f_{AAA}}{f_{AAA} + f_{AAG}}$$
$$f'_{AAG} = N_{Lys} \bullet \frac{f_{AAG}}{f_{AAA} + f_{AAG}} \quad (5)$$

where N_{Lys} is the number of triplets identical to lysine codons. However, such a formulation is not equally applicable to non-coding sequences.

1.3. Codon avoidance

Among UG- and CA-containing codons that tend to be increased by DNA methylation of CpG dinucleotides, five (AUG, CAA, CAC, CAU, and UUG) are generally avoided in coding sequences in vertebrate genomes, either caused by reduced amino acid usage or other unknown factors. Designating

these avoided UG- and CA-containing triplets as f_1 and the other UG- and CA-containing triplets as f_2, we define the triplet avoidance index as

$$I_{ta} = \frac{(f_2 - f_2') - (f_1 - f_1')}{f_1 + f_2} \quad (6)$$

1.4. Index of polypurine and polypyrimidine formation

Polypurine and polypyrimidine stretches are ubiquitous among eukaryotic genomes (Birnboim *et al.*, 1979; Mills *et al.*, 2002; Ohno *et al.*, 2002), but their frequencies in coding sequences are constrained by the necessity of codons with mixed purines and pyrimidines. For this and perhaps also other reasons, the polypurine and polypyrimidine triplets tend to be more frequent in non-coding sequences than coding sequences. We define the following index to measure the tendency of polypurine and polypyrimidine triplets:

$$I_{pp} = \frac{(f_{RRR,YYY} - f'_{RRR,YYY}) - (f_{Mixed} - f'_{Mixed})}{f_{RRR,YYY} + f_{Mixed}} \quad (7)$$

In this paper we demonstrate the utility of these indices in discriminating between introns and coding sequences.

2. Materials and Methods

The 10 annotated contigs from human chromosome 22 (ref_chr22.gbk), perhaps the best annotated human chromosome sequence, was retrieved from the FTP site of GenBank. The CDSs, exons and introns were extracted according to the sequence annotation in the FEATURES table, and their triplet/codon frequencies were computed, by using DAMBE (Xia, 2001; Xia and Xie, 2001). The indices shown in Eq. (1)-(7) were also computed by DAMBE for introns and CDSs.

For intron sequences, the indices differ little for the six different triplet frames (i.e., 3 on each strand), and the numerical results are presented only for the triplet frame starting with the first intron site.

The sequences are grouped into length categories. The indices from sequences with $L \geq 2000$, referred to hereafter as the training set, were used in a linear discriminant analysis performed with the DISCR procedure in SAS (SAS Institute Inc., 1989). The normal-theory methods (METHOD=NORMAL) was used, equal variance (POOL=YES) of the variables (indices) in the two groups (CDSs and intron) was assumed, and the prior probability was left as the default value of 0.5. Multivariate analysis of variance (MANOVA) was

performed to test the significance of the difference in these indices between the two groups. The fitted discriminant function derived from the training set was used to classify shorter coding and intron sequences grouped in length categories 1000-1999, 500-999 and 200-499 to investigate how the discriminating power would change with the sequence length.

To check the general utility of these indices in discriminating between coding and non-coding sequences, we have downloaded the Refseq file (zebrafish-rna.gbff) containing 6668 zebrafish (*Danio rerio*) protein-coding genes. The gene sequences contain only the CDS and their 5'-end and 3'-end flanking sequences. The coding sequences were again extracted by using DAMBE (Xia, 2001; Xia and Xie, 2001), and their indices similarly computed. The discriminant function derived from the training set was then applied to the classification of these sequences.

3. Results and Discussion

The annotated human chromosome 22 contains 111 CDS and 1824 introns with $L \geq 2000$. In this set of sequences, the MANOVA test shows that the five indices differ significantly between the CDS and intron sequences, with $F = 1770.72$, $DF_{Numerator} = 5$, $DF_{Denominator} = 1929$, $p < 0.0001$. Univariate significance tests (Table 1) show that all five indices can contribute significantly to the discrimination between CDS and intron sequences.

Table 1. Results of univariate significance tests.

Index	STD_T(1)	STD_P(1)	STD_B(1)	F	p
φ_{Nuc}	0.0565	0.0245	0.072	8340.66	<.0001
φ_{DiNuc}	0.0811	0.0392	0.1005	6364.42	<.0001
I_{pp}	0.0578	0.0526	0.0341	407.07	<.0001
I_m	0.0848	0.0838	0.0184	46.41	<.0001
I_{ta}	0.1079	0.0896	0.0849	868.1	<.0001

(1) Subscripts T, P and B stands for total, pooled and between, respectively.

The estimated parameters of the linear discriminant function for the training set are shown in Table 2, obtained from the linear discriminant analysis. The discriminant function can be used to classify unknown sequences. In short, the five indices are calculated for each sequence and Y_{CDS} and Y_{Intron} are then computed according to the estimated parameters in Table 2. A sequence with $Y_{CDS} > Y_{Intron}$ is classified as a CDS and otherwise as an intron. In this limited study, we used only CDS and intron sequences to demonstrate the discriminating power of the indices between these two classes of sequences. A practical study for gene-finding involving these content sensors would include many other classes of sequence states. For example, the

GENSCAN program (Burge and Karlin, 1997; Burge and Karlin, 1998) uses 17 different sequence states.

Table 2. Linear discriminant function

Index	Y_{CDS}	Y_{Intron}
Constant	-66.030880	-19.142580
φ_{Nuc}	334.618640	-53.342820
φ_{DiNuc}	65.782790	87.759840
I_{pp}	60.572250	61.689980
I_{m}	42.675790	43.121830
I_{ta}	-17.072500	6.354970

For these 1935 sequences in the training set, six coding sequences out of a total of 111 were misclassified as intron and two introns out of a total of 1824 were misclassified as coding, with the error rate of the classification being 0.0276 (Table 3). The discriminant function (Table 2) derived from this training set can be used successfully in discriminating between the CDS and the intron sequences, but the power of discrimination offered by these five indices decreases with decreasing sequence length (Table 3). It is noteworthy that the sequences "misclassified" with $L \geq 500$ are nearly always annotated in the GenBank file as hypothetical, and may have wrong annotation in the first place.

Table 3. Results of classification with the discriminant function

		Classified to		
L (bp)	From	CDS	Intron	Error
>2000	CDS	105	6	0.0276
	Intron	2	1822	
1000-1999	CDS	225	18	0.0376
	Intron	1	876	
500-999	CDS	155	23	0.0717
	Intron	10	696	
200-499	CDS	80	29	0.2494
	Intron	156	514	

I wish to illustrate the discriminating power of these indices by a particular "intron" that has its index values similar to coding sequences, and is classified by the linear discriminant function (Table 2) as a coding sequence with a posterior probability nearly 1. The "intron" belongs to a gene annotated as "LOC284861" in the ref.chr22.gbk file, starts with GT and ends with AG, and is "derived by automated computational analysis" according to the FEATURES table in the GenBank file. However, it is annotated as part of the coding sequence in other cloned homologous human cDNA sequences (GenBank accession: AL117481, AL122069, AL133561).

There are three lines of evidence to suggest that this “intron” is not an intron. First, when the intron and its two flanking exons are treated as a single exon, there is no embedded stop codon. Second, it has at least four indels when aligned with the GenBank sequence XM_375042, and all indels are inframe triplets. Such indel events are typical of coding sequences. Third and perhaps the most important, aligning the “intron” with other homologous human cDNA genes shows that its starting GT and ending AG are not conserved, which is not what we would expect if the starting GT and ending AG represent true donor and acceptor sites. All these suggest that the “misclassification” of the intron as a coding sequence by the discriminant function may in fact represent correct identification. When hypothetical sequences are excluded, then the error rate of the classification is decreases by nearly one order of magnitude.

As a test of the general utility of these five indices in discriminating coding and non-coding sequences, we have extracted the coding sequences from 6668 protein-coding genes in the zebrafish-rna.gbff file retrieved from GenBank. These coding sequences range in length between 117 bp and 18600 bp. Also extracted are the sequences upstream from the initiation AUG codon and downstream from the termination codon.

The application of the discrimination function (Table 2) to the classification of these three classes of zebrafish sequences (Table 4) shows that only a small fraction of the sequences were misclassified, i.e., when CDS sequences were not classified as CDS sequences and when upstream and downstream sequences were classified as CDS sequences. This suggests the similarity between intron and the upstream and downstream sequences. In other words, the discriminant function, which is based only on the difference between CDS and intron sequences, can not only pick up CDS sequences from a mixture of CDS and intron sequences as shown in Table 4, but it can also separate coding sequences from a variety of non-coding sequences.

Table 4. Classification of upstream and downstream sequences.

Sequence	L (bp)	N	To CDS	Error %[(1)]
CDS	≥500	5948	5454	8.305
	<500	720	667	7.361
Upstream	≥500	71	10	14.085
	<500	124	28	22.581
Downstream	≥500	3198	74	2.314
	<500	602	76	12.625

(1) Percentage of misclassification.

In short, the five indices we develop in this paper may be used for detecting protein-coding genes across all vertebrates. The test results are better if the discriminant function is applied to DNA sequences from other mammalian species (e.g., mouse and rat) instead of from the zebrafish.

We should finally mention four shortcomings of this study. First, the indices in Eq. (1)-(7) are still rather crude and can be much improved. For example, φ_{Nuc} and φ_{DiNuc} may have better statistical properties if they are based

on trinomial distributions. Second, there are other significant content sensors that can be derived from methylation patterns. For example, the ratio of UG-containing triplets and CA-containing triplet is nearly constant across the three triplet sites, i.e., (1,2), (2,3) and (3,1), in non-coding sequences, but differs dramatically in coding sequences. Third, the footprint of DNA methylation on non-intron and non-CDS sequences has not been thoroughly explored. The forth shortcoming is inherent in content sensors in that the indices are for detecting genes, but not for predicting the exact exon-intron boundaries. The latter would require information on signal sensors such as splicing sites (Foissac and Schiex, 2005; Gelfand *et al.*, 1996; Tenney *et al.*, 2004).

In summary, given the fact that even relatively crude formulation of these indices can allow us to discriminate between coding and non-coding sequences, we believe that the differential footprints on coding and non-coding sequences left by methylation-mediated substitutions can serve as powerful content sensors in gene detection in vertebrate genomes.

Acknowledgments

This study is supported by grants from University of Ottawa and from NSERC's discovery and strategic grants. I thank many of my Russian colleagues who have provided many helpful comments and suggestions. Two anonymous reviewers corrected errors in an early draft presented in the BGRS'04 conference. Two other reviewers corrected additional errors.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990, Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Baldi, P., and Brunak, S., 2001, *Bioinformatics: the machine learning approach.* The MIT Press, Cambridge, Massachusetts.

Bestor, T.H., and Coxon, A., 1993, The pros and cons of DNA methylation. *Curr. Biol.* 6: 384-386.

Birnboim, H.C., Sederoff, R.R., and Paterson, M.C., 1979, Distribution of polypyrimidine. polypurine segments in DNA from diverse organisms. *Eur J Biochem* 98: 301-307.

Borodovsky, M., and McIninch, J., 1993, Recognition of genes in DNA sequence with ambiguities. *Biosystems* 30: 161-171.

Brauch, H., Weirich, G., Brieger, J., Glavac, D., Rodl, H., Eichinger, M., Feurer, M., Weidt, E., Puranakanitstha, C., Neuhaus, C., Pomer, S., Brenner, W., Schirmacher, P., Storkel, S., Rotter, M., Masera, A., Gugeler, N., and Decker, H.J., 2000, VHL alterations in human clear

cell renal cell carcinoma: association with advanced tumor stage and a novel hot spot mutation. *Cancer Res* 60: 1942-1948.

Burge, C., and Karlin, S., 1997, Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.* 268: 78-94.

Burge, C.B., and Karlin, S., 1998, Finding the genes in genomic DNA. *Curr Opin Struct Biol* 8: 346-354.

Fatemi, M., Hermann, A., Pradhan, S., and Jeltsch, A., 2001, The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. *J. Mol. Biol.* 309: 1189-1199.

Foissac, S., and Schiex, T., 2005, Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* 6: 25.

Frishman, D., Mironov, A., Mewes, H.W., and Gelfand, M., 1998, Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 26: 2941-2947.

Gelfand, M.S., 1989, Statistical analysis of mammalian pre-mRNA splicing sites. *Nucleic Acids Res* 17: 6369-6382.

Gelfand, M.S., 1992, Statistical analysis and prediction of the exonic structure of human genes. *J Mol Evol* 35: 239-252.

Gelfand, M.S., Kozhukhin, C.G., and Pevzner, P.A., 1992, Extendable words in nucleotide sequences. *Comput Appl Biosci* 8: 129-135.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A., 1996, Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. U S A* 93: 9061-9066.

Hayes, W.S., and Borodovsky, M., 1998, How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.* 8: 1154-1171.

Meyer, I.M., and Durbin, R., 2004, Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res* 32: 776-783.

Mills, M., Lacroix, L., Arimondo, P.B., Leroy, J.L., Francois, J.C., Klump, H., and Mergny, J.L., 2002, Unusual DNA conformations: implications for telomeres. *Curr Med Chem Anti-Canc Agents* 2: 627-644.

Ohno, M., Fukagawa, T., Lee, J.S., and Ikemura, T., 2002, Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* 111: 201-213.

Pearson, W.R., and Lipman, D.J., 1988, Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.

Pevzner, P.A., Borodovsky, M., and Mironov, A.A., 1989, Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J Biomol Struct Dyn* 6: 1013-1026.

Pevzner, P.A., 2000, *Computational molecular biology: an algorithmic approach.* The MIT Press, Cambridge, Massachusetts.

Rideout, W.M.I., Coetzee, G.A., Olumi, A.F., and Jones, P.A., 1990, 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* 249: 1288-1290.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O., 1998, Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26: 544-548.

SAS Institute Inc., 1989, *SAS/STAT User's guide. Version 6, Volume1*. SAS Institute Inc., Cary, NC.

Sved, J., and Bird, A., 1990, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA.* 87: 4692-4696.

Tech, M., and Merkl, R., 2003, YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol* 3: 441-451.

Tenney, A.E., Brown, R.H., Vaske, C., Lodge, J.K., Doering, T.L., and Brent, M.R., 2004, Gene prediction and verification in a compact genome with numerous small introns. *Genome Res.* 14: 2330-2335.

Tomatsu, S., Orii, K.O., Bi, Y., Gutierrez, M.A., Nishioka, T., Yamaguchi, S., Kondo, N., Orii, T., Noguchi, A., and Sly, W.S., 2004, General implications for CpG hot spot mutations: methylation patterns of the human iduronate-2-sulfatase gene locus. *Hum Mutat* 23: 590-598.

Xia, X., Hafner, M.S., and Sudman, P.D., 1996, On transition bias in mitochondrial genes of pocket gophers. *J. Mol. Evol.* 43: 32-40.

Xia, X., 1998, The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Mol Biol Evol* 15: 336-344.

Xia, X., and Li, W.H., 1998, What amino acid properties affect protein evolution? *J Mol Evol* 47: 557-564.

Xia, X., 2001, *Data analysis in molecular biology and evolution.* Kluwer Academic Publishers, Boston.

Xia, X., and Xie, Z., 2001, DAMBE: Software package for data analysis in molecular biology and evolution. *J. Hered.* 92: 371-373.

Xia, X., 2003, DNA methylation and mycoplasma genomes. *J. Mol. Evol.* 57: S21-S28.

Xia, X., 2004, A peculiar codon usage pattern revealed after removing the effect of DNA methylation. In *Fourth International Conference on Bioinformatics of Genome Regulation and Structure*. Vol. 1 Novosibirsk, Russia: IC&G, Novosibirsk, pp. 216-220.