山形大学紀要（工学）第28巻　第２号　平成17年２月
Bull. Yamagata Univ. (Eng.) Vol.28 No.2 Feb. 2005

Genomic DNA sequences of non GT-AG introns in human mRNA genes

KUDO Yoshihiro, SAKAI Takamitsu, SATO Noriko, and Makoto Kinouchi

Bio-System Engineering, Faculty of Engineering

（平成16年10月４日受理）
e-mail（KUDO): chemies@yz.yamagata-u.ac.jp

Abstract

We searched human genome DNA sequences in the DDBJ/GenBank/EMBL for introns of mRNA genes which do not conform to the GT-AG rule, and collected 5791 fragments which do not form exon parts. Of these 159 are not of GT-AG form. Then we eliminated 19 because of non introns that were yielded by clerical error, frameshift, edition policy, and so on. Major part (94) of the 140 remaining sequences can be considered also to be GT-AG forms with alternative interpretation. There are several mRNAs carrying more than one intron where not GT-AG forms but non-GT-AG ones are chosen. This suggests easy usage of easy selection, even when there is more than one candidate, by easy computer software to infer an intron sequence as the logical difference between a gene and its corresponding cDNA.

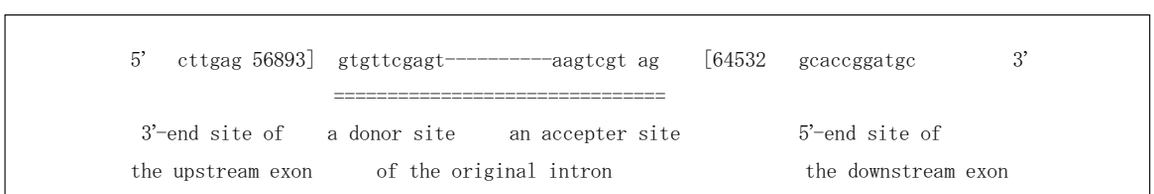
1. Introduction

Intron is a portion of genomic DNA (and then RNA) sequence which is transcribed to an RNA but is to be finally eliminated by splicing. Many introns begin with the dinucleotide GT (or GU) and end with the dinucleotide AG, and therefore are called the GT-AG (or GU-AG) introns, which are described as those conforming with the GT-AG rule.[1)] A sequence of an intron is determined by direct sequencing or inferred as a logical difference between an original gene and its corresponding cDNA:
“intron”= “upstream exonntron In order to know how a genomic DNA is constructed, we examined non GT-AG introns. In the present paper, we inquired relationship between human mRNA introns and their distribution (the ratio of occurrences). But as peptides to support our annotations, also those from other than human are used.

2. Method

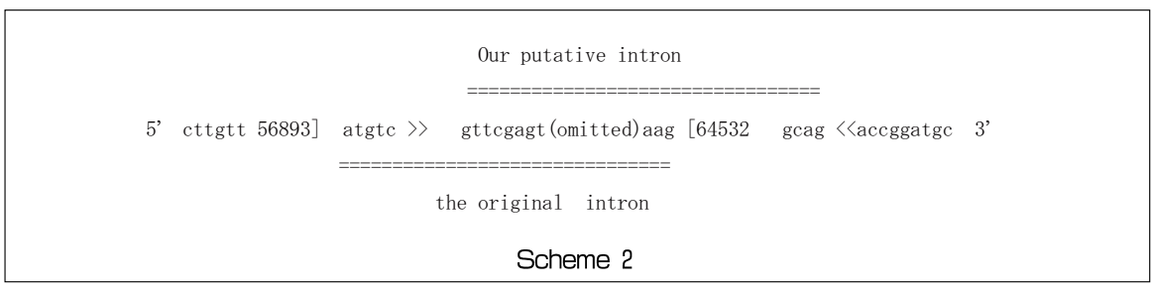
2.1 Notation

An intron is flanked by two exons, and have 5’(or donor) and 3’-(or accepter) ends. The position of the 3’-end of the upstream exon is shown with a subsequence and a distance from the 5’-end of the DNA fragment, and a sign “]”, and the position of the 5’-end of the downstream exon is shown with a sign “[“ and a distance from the DNA fragment and subsequence of the downstream exon (Scheme 1)



Scheme 1

An intron with our alternative annotations is shown a pair of "〉〉"(for a donor) and "〈〈"(for an accepter) (for example, Scheme 2)



Scheme 2

2.2 Procedure

We collect sequences of introns whose both ends (5'-/donor and 3'-/acceptoer) are described specifically, and eliminate those which are described to be putative/hypothetical ones. First of all sequences which may be not introns are eliminated. Then we try to interpreat as many introns as GT-AG ones as possible

3. Results

We searched the DDBJ/GenBank/EMBL, Release 34 for DNA sequences of human mRNA introns, and those of other mRNA introns to support our alternative annotations, excluding putative and/or hypothetical ones. Then we searched three databases: SWISS-PROT (SW), PIR, and PRF for the same peptide sequences as those of peptides to be finally produced if the alternative annotations be true. We accessed all of the databases also through GenomeNet managed by Kyoto University (http://www.genome.jp/). Among the 5791 sequences examined there were 159 of non GT-AG form as shown in Tables 1 to 9.

A pair of a donor and an accepter subsequence of each intron is separated by its Accetion number assigned by the DDBJ/GenBank/EMBL and displayed with subsequences of the upstream and downstream exons, and its identification number (of the SWISS PROT, PRF or PIR database) of one of the peptides to support our alternative annotations is attached (with six exceptions [a GT-AG and four GT-TG introns in Table 6, and the D63805 intron in Table 9).

Table 1 lists sequences that are clearly not and/or maybe not introns.
Table 2 shows thirty five cases where our alternative annotations lead to the GT-AG forms without changing sequences to be matured mRNAs.
Table 3 shows eight cases that our alternative annotations change introns to the GT-AG forms without changing peptides to be produced because of synonymous codons.
Table 4 enumerates forty six cases that our alternative annotations to be GT-AG introns lead to different peptides from those of the original annotations.
Table 5 is a group that conversions to GT-AG introns need reinforcement by insertion of one or more nucleotides.
Table 6 collects eleven GT-GG, four GT-TG, and three GT-CG introns. The three GT-CG introns have no support to date.
Table 7 shows nineteen GC-AG introns.
Table 8 is a list of one GA-AG and two GG-TG introns.
Table 9 shows two other introns. One of them has no support.

Table 1. Sequences that may be GT-AG introns or that may be not introns (19)

Upstream EXON	Donor	ACCESSION No.	Acceptor	Downstream EXON	Remarks
>><<gccctatggt 1095]	cggccgggccggggcctgcgg	L41919	atgaagcacgagccgggcctgggta	[1597 gctatggcgac	new data
><<ccccggactc 3459]	gctgtcggaccgcgacagcta	U07802	cggaccgcgacagctacctaagcgg	[3490 ctcc	SW:P47974
ggagctgaaa 11453]>>	gtgagtgaaaatggagggcaa	U18671	tgtgactcctgtcttctccccaca(c)G	<<[11620 acggacaccct	PRF:2108377A
ctctaataga 14753]>>	gtgagatatgaactgttcatt	U18671	ctttgtctgtctttgggtatccag(g)	<<[16764 caggtggatga	PRF:2108377
cctggagag>> g 133]	tgaggctcc V00489 cag		<<gatgttcctgtccttccccaccaccaagacctacttcccg	[290 cacttcg	SW:P01922
aacttcaag 454]>>	gtgagcggcgggccgggagcg	V00489	actgaccctcttctctgcacag	<<ctc [598 ctaagccactg	SW:P01922
ctccctgagct caggctgcag3428]>>	gtaagatgaaggaggctgatc	X17093	aggtcctgtttttgttctacccca(a)G	<<[3564tcactgacagt	SW:P30511
cgattcccag 1121]>>	gtacggcccggcctgaccctg	X52601	agtggctgttgtttgatattctctc	[1437 cttttctttttccag <<at	SW:P11387
cagctgtgtg(a) 1470]>>	gtatcaaccctgctgccctg	X53682	aatctgtctagcccttgttctacag	<<[2043attccaaacca	SW:P13236
>><<ggaggcgacc 1330]	gtgagtgggacgggcagggac	X73637	cag		
	<<gtgagttgctacagtggctggagagacgacatctgctccatgggctggtggccgacagtaatctcacgctgggacctggcctgcagcccc			[2514 tgccc	SW:P51460
atcctttatg 410]>>	gtaactacannnnntctttca	X80763	gtaactacannnnntctttcag	<<att [436 atgtctggcca	SW:P28335
cggccgtcag(t) 415>>]	gtaggccggcgtctgcggga	X96924	cctctcacccgcctcccgcctccag	<<[484 gtttggaatgt	SW:53007
ccggatgcag 1764]>>	gtggtggggccgggaaggggga	X96924	gcccctgcctgctgatgtccccata G	<<[1846 ggcctggaggc	SW:53007
>><< aagtgggaag 24831]	aaagactgggatgatgaccaa	D42052	agactgggatgatgaccaaaatgat	[24859 tgaaattggct	SW:P50900
gtggtgctcc 5883]>>	t	D89870	<<[5885gatgcccctca		SW:P54368
gccgccggca 1268]>>	cc	M34677	<<[1271ccagcccgggc		SW:P23610
ctaacttaag 29272]>>	c	U81031	<<[29274 tgcgtggactg		SW:Q99490
caaggaccag 705]	aaggctggaaggggtctgtgc	X75015	catggctgtgtccttttttttcag	[869 attgtggacct	SW:P16050
attcattcct 83]	nnnn	X80763	[88 ntttttcag <<tgtgcac		SW:P28335

Table 2. Introns that can be converted to GT-AG ones without change resulting exons (35)

aagccaaa 530] g>>	gtgggacccacggggtgcga	AJ001563	cgctgtgccaacctctgtccctaca	[628 g	<<ggcagccccg
catgccca 82] g>>	gtaagccaacccaggcctcg	AJ001563	cgcatccacctccatctcttcctca	[201 g	<<cacctgagtt
aagccaaa330] g>>	gtgggacccacggggtgcga	AJ001564	cgctgtgccaacctctgtccctaca	[428 g	<<ggcagccccg
aaaaaag>> g5395]	taagggctcatcctctgtgaa	AJ001981	actctcttgcttctctttacacag <<g	[5483ctggaaaaatg	
ccccacg>> g5124]	tatgtaatgccttatgggctg	AJ001981	tcccttctctgtgttctcacccag <<g	[5233 cagtgtttatg	
gccaggg>> g5563]	taaatagtcctttcaggccaa	AJ001981	atttgttttctcttctcccctcag <<g	[5649 tcctttacgac	
cttcattat6625] g>>	gtaagtagagtcaaaatgac	AL022069	aaaatgcgggtctctgtcccccaca	[8620 g	<<cctttcagac
acgcccaca57453]g>>	gtgtgtgccacagtggttcc	AL022069	ctgaactctgcccctgtcacctaca	[65669 g	<<actctcctgg aga
tcaca 16015] g>>	gtttgtaaaagacg	AL022069	tgtgctgtgtttgcatgta	[18252 g	<<atttcggcct
ccaagtc156] g>>	gtaagtgaggagggccacccL15346		cacctcccatctccctcctctcca	[261 g	<<gcactacctg
attccctca386] g>>	gtgtgtgagtgtgtcctggg	L15346	cagaccttccctgacccactgctca	[476 g	<<ggatggagac
cagatgacg2373] g>>	gtgagggcggctcgcc	L19314	cccgtctgtctctttctggcccgca	[3006 g	<<ctgcgctgag
atcgg>> gta 1352]	agagtgaatgaytt	U90269	tgttattgttttccctttctag	<<gta	[1830 gataa
tggagag>> g 230]	tgaggctccctcccctgctcc	V00488	ccctcactctgcttctccccgcag	<<g	[348 atgttcctgtc
gtttccca184] g>>	gtaaggatggagaacactag	V00558	atttgtttttatgtttccaatctca	[310 g	<<gtgccagatg
tacttca>> g 5804]	taagtagcagggaggggcttc	X03562	ctctctgtgctgtgggacttccag	<<g	[7506 caggcccgcaa
gagatcacc612] g>>	gtgagccccccctgctcctgc	X03900	gtctgatttccctcttcccccaaa	[1336 g	<<gcaagcacga
aaatacac>> g 1401]	tgagtcctggcgcca	X03900	tgcctttcctctctgcacgtccag	<<g	[1520 ctgccccccgg
tttccgca219] g>>	gtgagtctgggctgcgggag	X63863	ggccctcacctgctgtcactcccca	[335 g	<<accccaacgt
actttaaa1257] g>>	gtaagactttcagatgtcaa	X64171	cacctctcttttctttgtattttaa	[2308 g	<<cccagtgggc
cttttcca727] g>>	gtaaaactttcttaaatta	X67336	ttctctctttctcctttgttttta	[1413 g	<<ctggacagac
agtgacg>> g 463]	tgaggaggggtgggggtagg	X70082	gcagctctctgctttgccctgcag <<g	[576	ctctggctggg
taaaag>> gt 571]	gattcatggggaactagagat	X70197	ccagggtgtctctgtgtttgcag <<gt	[675	gtccagtgtga
taaaag>> gt 250]	gattcatggggaactagagat	X70203	ccagggtgtctctgtgtttgcag <<gt	[354	gtccagtgtga
taaaag>> gt 250]	gattcatggggaactagagat	X70204	ccagggtgtctctgtgtttgcag <<gt	[354	gtccagtgtga
taaaag>> gt 250]	gattcatggggaactagagat	X70205	ccagggtgtctctgtgtttgcag <<gt	[354	gtccagtgtga
taaaag>> gt 250]	gattcatggggaactagagat	X70206	ccagggtgtctctgtgtttgcag <<gt	[354	atccaatgtga
taaaag>> gt 571]	gattcatggggaattagagat	X70207	ccagggtgtctctgtgtttgcag <<gt	[675	gtccagtgtga
taaaag>> gt 339]	gattcatggagaactagagat	X70208	gaccaatgtctctgtgtttgcag <<gt	[441	gtccagtgtga
catgcaa>> g 2201]	taagtggacttcattgttccc	X75598	ctcattctctgtcctgttgaatag <<g	[6404	cttccgaagat
gccatg>> gt 6506]	gagtgtgcctgtgtgggatac	X75598	catatcttcttctgtccttggag <<gt	[7591	ctgggaggggc
tctaatgg 10739]	g>> gtgagttttccaacagttgc	X91233	atttaatctctctctatattttgca	[13493 g	<<aatgtaacag
tggtggaa 26818] g>>	gtaagcaagacgtg	Y09912	attacctttactgctgtctttttca	[28883 g	<<gagaagctgt
ttaaaaaa 17406] g>>	gtatggataattccccccaa	Y09912	tatggcaatttttttctctctttcta	[24530 g	<<ttccagttcc
cttcaaca 13503] g>>	gtgagccactgctttctgca	Z83307	tggctatttctttcatccacttcta	[14194 g	<<gactcatttc

Table 3. Introns that can be converted to GT-AG ones without change of peptides after translation.(8)

gctggcagg>>	g 97022]	taagcagtcacctcaaacagg	AL022069	ccacaaatttgtctaattatacag	<<a	[98562 tttacccctttt
catcctcaa>>	g 26233]	tgagtagtgtgggagccaccc	AL022069	ctaaaataggatcttgtttttcag	<<a	[27983 gccaagctggg
atggaccaaa11094]	tg>>	gtaatttctaaattaatat	D42052	tttctttttttaactgttacttttg	[11839	ag <<gaatgaaca
cctgggcag>>	g 265]	ttggtatcaaggttataagag	V00505	cttgggctgttttcctaccctcag	<<a	[394 ttactggtggt
cttgggcag>>	g 2171]	taagcattggttctcaatgca	V00508	atctgatatggtgtcatttcatag	<<a	[2294 ctcctcgttgt
cctgacgg.>>	gt 421]	gagtgccctgcggccgcgggg	X63863	tgaggggaggtccccccaacag	<<gc	[577 gacctggagag
tttccgg>>	gta 654]	agctacaggctgctgtggcca	X63863	atcctcgcttcgggtcctgcag	<<gtt	[748 aaccg
cctcagaag>>	g 24751]	taaccccaccacgaaaaacaa	Y09912	gaatgttaattcttgcaatttcag	<<a	[26702 gccaaatcgaa

Table 4. Introns that can be converted to GT-AG ones affording different peptides from the original annotations (47)

ccacggacag 38998]>>	gtaacaggcct	AB008681	tcacatag	<<gtaggcacga	[39741	gacggtacatg
gcaaggctgc39864]agacg>>	gtaagtaggatggcag	AB008681	ctgccaggctctctctttctcctag	<<[39963	gacccgtggat	
aaaatc>>	gtaa24575]gtatagtgtgtaattgtgatt	AC004237	ttaaaggatttcgtttatttattaa[28194	ag	<<gttctgttg	
tagag>>	gtaag19057]gataattgaaattggtggagt	AC004237	taatttttttttag	<<gttaaagaaag	[24477	acctgatgtgg
gtacttgaag10966]	gctacaatg>>	gtaagtgaatct	AC004237	aactgattctcttcatcattttcag	<<[18682	ggactattttt
gggaaggacg1240]>>	gtgagtgtccacgccctcctc	AF020276	ctcctcctgtgtgtgtatatctcct	[1472	ag	<<gacagaatt
caggctgcgt2579]>>	gtaagtgatggcggcgggcgt	AJ001977	ggtctttttttttgttctaccccag	<<g	[2687	cagcaacagtg
cttgtaaag>>	g 2734]tgagattctggggagctgaag	AJ001977	tgttcatgactattgtgttctgtag	<<[2898	cctgagacagc	
agcaccacac132]>>	gtgagtgcctcgggagaggag	AJ006854	gcctgcctacctgcttctccttcta	[266	g	<<gaaaagctga
ggggatcagt2735]	gtaagtccagtttcaacctgc	AJ224869	atgcttgctgaattggaagtgaatg	[4844	tccattcctttgcctcttttgcag	<<atata
ccggctgg>>	gt28072]aagattccccgggacacccca	AL022069	gatgatcatttctgtgttgttgaca	[46997	g	<<gtgctggcat
gaagcctgag12395]aa>>	gtaagtgagaaaaaactag	AL022069	catctgttttctctccctttctttt	[15977	ag	<<catcctcct
cgtgctgcgg21888]>>	gtgcggaggggcgggccgggc	D26607	tctgatccactgtgctcttttccga	[21975	cag	<<gatcagca
gaaagaa>>	gtg2456]	agtccgggcagcgccttcccc	D43639	aatcgggtctgcttgtgttttccag	<<[2604	gtggaataagt
tctggccccg1337]	gatggtacggcgccaacccagacccgcgcttccccgcca>>	gt	D50739	gagcctgtgcgtttgtcgcttacag	<<[2329	tctcccgcttc

SW:Q13705	
SW:Q13705	
SW:O15066	
SW:O15066	
SW:O15066	
PRF:2401341A	
PRF:2314363	
PRF:2314363	
SW:P19440	
PRF:2413277A	
SW:Q15349	
SW:Q15349	
SW:P29474	
PIR:JN0684	
SW:P43699	

（Table 4）

gaggcagagg4662]acctgcagg>>	gtgagccaaccg	L15440	tgctctgcgcggcacgtcctggcag	<<[5458tggggcaggtg		SW:P01308
cagctacag>>	g279] taaggggcttcctagtcctaa	L22582	ttattcacccctgtgtcctctccac	[363 ag	<<gtgtccagt	PIR:S49530
tcaggaa>>	gtg1725] agtaggggcctggggtctggg	M24543	attcctgcgtctgcttcctccccag	<<[3351caaaagcgtga		SW:P07288
cgag349] acttccgcaccgtcctggttaacgcggatggcgag>>	gtgggt	M94363	tgccgccgtgtcctgtcccgcacag	<<[508 gaagtggccat		PIR:45023
atcgaagag>>	g 2836] tgagcgctgcactccctccct	U00921	catcag	<<[3593caccttctgtcctggtccc([3612Error]aggcccagggc		PRF:2402239A
ccg 614]>>	gtgagtgtggttgcgtgtgtg	U02948	catgcctgtctccttgcctgcccca [1319	ccgtgtcag <<ga		SW:P11686
aacac 1477]	acggagatg>> gtgagaggtgtg	U02948	ggggaagaccaggtggctccatgcc	[1823 tttccccag	<<gt	SW:P11686
gctaaggtgg1963]	>> gtatttctggccacagatgca	U19759	cgttccatgttcatctgacttcagg	[2817 tag	<<agctacag	SW:Q12986
tgcacttca>>	g16575] tatgaagcaaaccggagaggc	U24578	ccacttttgcgcacgtcatccccgc	[16835 ag	<<ccaaggatg	PRF:1006211A
gagggaca>>	gt16971] gagtcatctggtcccctcagt	U24578	tgttgactgtccctttgnccctgc	[17111 agnccccag	<<ta	PRF:1006211A
accacag>>	gta857] acagtgatggcttcatttcnt	U25441	ccttccccctctcatccacag	<<ggg [907 accccactgtc		PRF:1705199A
gtgcttg>>	gta1386] agtttgggttggcttgagctg	U25441	tctctcttatttggcaacttag	<<ggg [1445	ccttcattgtc	PRF:1705199A
agctttgg>>	gt 1605] aagagacaccagcattgcaga	U33317	ctgatgctctctctctgtcccca	[2517 g	<<gctcaacaag	SW:901524
cttcaggtga 9234]	gtagcaaaggacaccaaggaa	U38291	atcagacatatcttatctcccttct	[9427 aggtgactctg		PIR:A43359
gg>>	gtgagcca11256] gctcccggatgagtgaaccaa	U70065	agatcctgccttctccag	<<gccgcca	[12009 gagctgcggctgat	SW:P52333
gg>>	gtgaa26176] gccctgggtgagtaaggttta	U81031	actgaggtttcatcttcttccgaca	[26359 g	<<aagccactac	SW:Q99490
atgacccactgaggattacaagaagct>>	gtgagtatgcaacgccag 54125] gctca	U82671	ctctcccctctcttag	<<g [54937 gcaccctacaa		SW:O15231
gaggccagcg 720]>>	gtgcgacccccgtccttcccc	U83668	catgccctcccgccccacgcag	<<gtg [802 gcaggtacgtg		SW:P05217
agaagacaa>>	g1943] tactgtttcggttttactcca	U90269	ctgtcctgttgtggccctcaaaatc	[2142 gaacgacag	<<tggtg	PRF:2401351A
caggctgcgt3041]>>	gtaagtgatgggggtgggagt	X03945	ggtcctgtttttgttctactccag	<<g [3149	cagcgacagtg	SW:P10317
ccacttacct438]>>	gtaagggctgggggcattttt	X58899	tccagcccccaactcctcctgcag	<<a [722 caagctggtgt		SW:P08686
caccctgag>>	g1825] tgcgtcctggggacaagcaaa	X58907	aggactccacccgatcattccccag	<<[2025attcagcagcg		SW:P08686
gccacttacc437]	t>> gtaagggctgggggcatttt	X58907	ctccagcccccacctcctcctgcag	<<[721 acaagctggtg		SW:P08686
aggcgtcaag807]	a>> gtgagtggggccccgggcag	X63863	tgaccgccctcttccttgtgccgca	[905 g	<<ttgacaagacggccaccaa>>	SW:P52656
g29026]	tgagtttattagactctgggc	Y09912	tctgtgcctcgcccacccctttgca	[31821 g	<<gcaactttgt	PRF:2218269A
gaagaagaag81]>>	gtggggaggggcaggggaggc	Y13586	ccaatgacccatttctggctcctca	[238 g	<<atgaggaaga	PRF:1505244A
agcgaggccg270]>>	gtgagtgaccccggcccgggg	Y14582	ggggctgaccgcggggccggggcca	[514 g	<<ggtctcacac	PIR:I37515
tctg>>	gtaagt6281] taatgtaaactcaaggaatat	Z73359	ggctacgttttcatttttttatcag	<<[7393atgtcttctcc		PIR:G02334
ctgtgtgact4627] cag>>	gtatactacagcattttt	Z96810	gaagtaattaacatttttttggtag	<<[5165gctggaattta		SW:P48067

(Table 4)

tggatttatg5264]>>gtaaatagtaactataaaatt Z96810 ctggaagaaaatattag <<caagcact [5521 caatgtactct SW:P48067

tggg2237]>> gtaag Z97370 cccag <<agccatcttcccagcccaccgtccccatcgtggggcatcattgctggcctggttctccttg [2400 gagctgt PRF:1311286A

cttcatg>> gta 1353] aggagtttaacattgaatatg X07963 ctttttaattataggtcctcag <<tgt [1483 gacatccagat

Table 5. Introns that can be converted in to GT-AG forms with reinforcement with nuceotides

A capital letter is a nucleotide to be knserted.

ttgaaaaatt 1440]>> gtaagttttt	AB010084 acctaattaaacaaattt	CTAATCATTTCTTATAG <<[1569	cttttactaaa SW:Q31241
catcccaaag3291]>> gtaaagtaaag	AF026029 agagatacaatgacaattcttttca G	<<[3772	ggtttgcgtat PRF:2201493A
aggGttacag>> g 9909]	tgggagtgcccttgtagtcc	U24578 cctccaccctctccctcccatgtag	<<[10045 cctmagatcca SW:01028
ctgctacaag 658]>> G taagcactgctg	U41163	gcccctccaccccctccag	<<[749 gacgccatcat PIR:G02277
cgacaaccag1303]>> Gtttgcatgggctctgggacag	U41163 tcggcctgagctagcctggccacag		<<[1414tttgtaggtgt PIR:G02277
atctcttcag9452]>> gtgcccgctgggacgggttgg	U70065 cactgcccatgcccccaccccacga G		<<[10839actatgagctc SW:P52333
ggacagcatt15063] cGcgag>> gtaagagaaaggtcag	U81031 ctcttactgcccttctcccgttaca	[21547 gag	<<gctgtgat SW:Q99490

Table 6 GT-XG introns (4 GT-CG, 11 GT-GG, and 4 GT-TG ones)

ctccacag9615]>>	gtctagaggccaggcaggaac	U18671 ttctctccacacctctctctcctcg	<<[9716 agcctttgtgg PRF:2108377A
ccccgctgcg4187]>>	gtaggcggcggggacca	U28054 cccgacctatctcggtccatctacg	<<[4389 ggtccgaggca SW:P26927
acaagccgca5006]>>	gtgagtccctggtgctcccgg	U28054 cgattttgctcccgctcccgcctcg	<<[5134 gttcacgttta SW:P26927
ataccgtggg 3309] gg>>	gtaagaaaggcccctgcag	U48705 tcctcttccttggtcccctcttctc	[3841 cg <<gactgcagt SW:Q08345

(Table 6)

aagtctaaca9291]>>	gtatctatctgtccctgggct	L24038	cctgtagtggtccttgacctggcgg	<<[9797	acatcttccta	SW:P10398
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48213	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48214	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48215	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48216	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48217	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48220	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48221	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48931	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
ccctgggcag308]>>	gttggtatcaaggttacaaga	L48932	ggcactgactctctctgcctattgg	<<[420	tctattttccc	PRF:070522A
aagtctaaca9291]>>	gtatctatctgtccctgggct	U01337	cctgtagtggtccttgacctggcgg	<<[9797	acatcttccta	SW:P10398

aggctgag>> gt68]	aagtgaatgcttgagcccagg	AF039597	tctttccttttctctctctccctag	[186	ctctg	<<gatcgt
ctcctcag>> gt63]	gagccctcacaacctctctcc	U42588	tctggcattgttgtcacaatgtgac	[414	aactg	<<gttcga
ctcctcag>> gt63]	gagccctcacaacctctctcc	U42589	cctggcattgttgtcacaatgtgac	[414	aactg	<< gttcga
ctcctcag>> gt63]	gagtcctcacaatttctctcc	U42591	cctggcattgttgtcacaatgtgac	[414	aactg	<<gttcga

Table 7. GC-AG introns (20)

acagtggaag14818]>>	gcaagtaaatgaaat	AF001295	ccttaccatgtgtctag	<<[	16913	gttacagtgac		SW:P17643
ccacagcaag4024]>>	gcaagtagaggga	AF019084	tcttctccccatatcag	<<[	4342	tatgaggagct		SW:P35908
tctgaagaag2522]>>	gcaagtgacaca	AJ000263	actccccacctttccag	<<[	3282	gatgtggactg		SW:P35908
agatgaactg620]>>	gcatgtgctgagtc	AJ000673	cttttacttgaagcag	<<[	2415	gattttatgag		SW:Q13241
agtttcagag13787]>>	gcaagtgttcattt	D42052	aatgttgtttatgttttag	<<[	14272	gtcatagaagg		SW:P50900
gggaccccag30192]>>	gcaagttctgcc	L23982	atctgtgtttctcacag	<<[	30441	ggtcaacctgg		SW:Q02388
tttggatcag2163]>>	gcaacctgcctccc	L46590	atcccattcttccacag	<<[	2259	taatgggggcc		SW:A54183
cagaagaaag3315]>>	gcgagtgggggtggagagggg	U28054	ttgatcttctgtctcttgctcccag	<<[	3393	actacatacgg		SW:P26927
cagaagaaag5243]>>	gcaagtgggggtggagagggg	U37055	ttgatcttctgtctcttgctcccag	<<[	5321	actacgtacgg		SW:P26927
gaaacaccag1017]>>	gcaagtgctcttctttatgct	U51038	gtttcgatgtcattttaaactacag	<<[	2428	agtggatccct		PIR:A47297
catgatgaag118086]>>	gcaagtgttactcagccca	U82828	ttttgtgtctttttttaatggta	<<[	119109	agagacggaat		PIR:A4310
ctgccccaag44163]>>	gcaagtgaccacaaatctgcc	U89335	tctcccctcccctccacag	<<[	45334	ggttttgaagg		SW:P31695
tgctactcag2914]>>	gcatgtgcccaccccttcccca	X56494	atgctctgtcccctcccag	<<[	3162	atgctggagag		SW:14618
gggttcggag2679]>>	gcaagtccccgttgtccctgg	X56494	cccccactttgtccatcag	<<[	2764	gtttgatgaaa		SW:14618
ttctgttaag212]>>	gcaagtaccagatgattccta	X94208	ttttcttggatttctttag	<<[	553	tatcgagaaaa		SW:P12270
tacatcgcag937]>>	gcgagtgcccagtggccgcat	Y00486	ctgtcacttaccctgacag	<<[	1931	gcctagactcc		SW:P07741
tctgaagaag2545]>>	gcaagtgacacaggattgaga	Y13621	acactccccacctttccag	<<[	3303	gatgtggactg		SW:P35908
caaggccaag22664]>>	gcaggctctgctcggcctccc	Z20656	cctgtcttgcttcctgaag	<<[	23001	gcaaacctgga		SW:P13533
tccaggcaat2098]>>	gctgagtaactggcggttgca	AF026029	aaattatttttttcctgata	G	<<[	2808	gctggcccggt	PRF:2201493A

Table 8. GG and GA introns

GG-TG (2)										
tgagaatttc357]	c>>	ggtaagaagaaaaatagatg	X91233	catttgtctatgattatatttttca	[1351	g	<<aaaccacatt			SW:P40933
ttcattttgg 1448]		g>> ggtaattttatctttaggca	X91233	gctttgctctatggtttttctttca	[2807	g	<<ctgtttcagt			SW:P40933
GA-AG (1)										
gacccccag5397]>>		gataggagttgggccagttatU28054		tcatccagcccaccccatcctacag	<<[5486		accaggtgcag			SW:P26927

Table 9. Other introns

Others (2)										
cgatccc>>	aat 1076]	aatgtctcctcccttaaatcc	AF010258	ccagtgtttctcttccccgcag	<<aat[2094		aacccagcagc			PIR:JC6553
agactcg atg 1573] >>		agttaaaatgaga	D63808	tacctccttcctcctcctgtagatgatg	<< [1721		tcaatgaggag			

4. Consideration

In order to make our alternative annotations reasonable and reliable, we used peptide sequences retrieved from the peptide databases as much as possible. However, there are many cases where one peptide is not always given only one sequence as described in some databases e.g. in terms of not only "VARIATION" but also "CONFLICT". By the way , a serious situation is that sometimes a peptide sequence is literally translated from a DNA sequence though there may be genetic complicated aspects, such as editing, frameshifts, and alternative splicing. Very short sequences may be not introns but may show frameshifts in translation to peptides. PIR I52571 (human glycophorin MiI; part) was translated from M81826 of "GB/EMBL/DDBJ" on the basis of the paper by Huang, C. [Blood (1992) 80,257], which was created on 20-JUL-1992 (Rel.32), and updated three times (Last updated, Version 4, Rel 60). Its sequence was revised on 02-JUL 1996. Three references are recorded: [1] Huang, C.-H, Spruell P., Moulds J.J., Blumenfeld O.O, Blood (1992) 80, 257, [2] Blumenfeld O.O, Submitted (20-JUL-1992) to the Database, and [3] Blumenfeld O.O, Submitted (27-JUL-1998). Update for Version 3 was done on (28-JUL-1998). [3] of the EMBL Version informs "Amino acid sequence updated by submitter" (its corresponding GenBank Version does not). In the peptide sequences, the first 26-common amino acid-subsequence "(1) LSTTEVAMHT STSSSVTKSY ISSQTN (26)" is followed by "(27) ICTN GTHMQPLLEL----" in PIR I52571 and "(27) DMHKRDTYAATPRAH-----" in EMBL M81826. The current DNA sequence, containing an GT-AG intron from 80-785, is corresponding to the amino acid sequence of EMBL81826 (Scheme 3).

```
      aat78    g79  80 gttt ----------tgcag785  786at   atg cac aaa cgg   ---
26 N(aat78)D(g79/ 80 gttt---intron-- tgcag785/786at) M(atg)H   K   R  ----.
```

Scheme 3

On the other hand the ICTN--- of PIR I52571 means that the gene using a non GT-AG intron (gg-ag) expresses (Scheme 4).

```
 aat 78       79ggtttgtttt ---tgcag785     786 ata    tgc aca aac gg---
26 N(aat78)  /79ggtt---  intron --ag785/   I(786ata)  C   T   N   G----.
```

Scheme 4

Sequences shown in Tables 2 to 5 can be considered also to conform to the GT-AG rule.

The intron of X73637 is a GT-CC one (Scheme 5),

```
--ggcgacc1330]    1331gt---cag2423
2424gtgagttgctacagtggctggagagacgacatctgctccatgggctggtggccgacagtaatctcacgctgggacctggcctgcagcccc2513
 [2514  tgccccagacc----
 (While the original intron  is 1331gt to cc2513, our putative intron is 1331gt  to  ag2423)
```

Scheme 5

which is to lead to a subsequence of a peptide GDR (c 1330/intron/2514 tg) PQT. On the other hand, the peptide shown in Reference No. 1 of X73637 has a subsequence (Scheme 6),

```
GDR(c1330       /1331gt---(intron)---  ag2423/          2424gt)
      ELLQWELLQWLERRHLLHGLVADSNLTLGPGLQP
      L(c2513         2514tg)PQT---..
```

Scheme 6

which comes from the same alternative annotation as ours to result in the GT-AG intron (See Scheme 5).

Table 2 is sequences with simple right or left shift of one (g), two (gt), or three (gta;U90269) nucleotides to make the GT-AG introns without any effects on exon portions.
Table 3 is a kind of silent shifts where changes of introns do not give effects on peptides to be produced.
Table 4 shows shifts to change peptides to be produced, but the original annotations themselves do not show peptide sequences. The code numbers of newly expected peptides are shown in Remark.

Sequences in Table 5 are a little modified to be changed to the GT-AG ones.

The first entry of Table 5, the AB010084, is repaired with "CTAATCATTTCTTATAG" by comparing with the sequences Y14287 and Y14289 which yielded an identical peptide SW:Q31241.

Table 6 collects the GT-XG introns other than the GT-AG ones. Although GT-TG introns have no supports, their figures may be reasonable.

Today, most of introns are inferred not only by direct sequencing but also by logical subtraction of cDNAs from DNAs or in black boxes called computer programs. As a result, when there is more than one possibility, more than one intron can be inferred. For example, a combination of a DNA XggtYaggZ and cDNA XggZ is to afford a set of three answers: tYagg, gtYag, and ggtYa. Among them, only gtYag conforms to the GT-AG rule. Easy computer programs may get and show only the first candidate (in this case, unfortunately a non GT-AG one), and stop. Easy users, taking account of only peptide sequences, may accept it without any discussion. For example, the AL022069 sequence contains seven introns, and the annotations of them adopted only the non-GT-AG forms (Scheme 7) (See also Tables.)

①	a 6625] g >>gt------ca [8620, g<<;	Simple shift
②	g 12369] aa>>gta---tt[15977 ag<< c;	Change from Ser(agc) to Asn(aac)
③	a 16015]g>> gt—ta[18252 g<<;	Simple shift
④	aa>> g 26233] tg---cag<< a[27983;	Silent change of K(aaa) to K(aak)
⑤	g>> gt 28072] aa—ca [46997 g<<gt;	Change of GlyGly(ggtggt)to Gly(ggt)
⑥	a 57453] g >>gt—ca[65669 g <<;	Simple shift
⑦	gg>>g 97022] ta---cag << a [98562	Silent change of G(ggg) to G(gga)

Scheme 7

The four introns of X63863 are similar to them. The annotation of X17093 that a computer software produced was adopted, though the user abandoned it[4] because of nonconformity to the GT-AG rule. In our opinion the user of this case, is not always reasonable. In general who should decide such a bold choice of a particular one of the alternative annotations? The international and useful databases in their early stage collected data obtained directly from real peptides, but recently at least one of them get in silico peptide sequences translated from the DNA sequences of the DDBJ/GenBank/EMBL frequently without any explanatory notes[3].

Since human always makes mistakes, we cannot avoid that wrong data slip among a huge amount of good data in our useful databases. It is necessary to establish better procedures for prevention of mistakes, easy detection of wrong data, effective watching of the databases and so on. Because in this field many new kinds of data continue to occur, frequently we meet suspicious pieces of information. However sometimes we cannot judge a reason why a data appears to be suspicious because of our ignorance or wrong data. Individually we try to ask the original authors as much as possible, and get good solutions. The first mail was sent to DDBJ in 1996, and the answer of one of them came from an EMBL person in London. However such nice pairs of questions and solutions do not become international and common knowledge among people who want them. So we need a cooperative system to inform each other whether unfamiliar data suggest new genetic aspects or only reflect any kinds of error.

Although our present cases themselves may have been corrected already or will be revised near future, worth of suggestions issued by the present paper would not decrease. The present paper tried to regard as many introns as possible as GT-AG ones. But conversely all of the apparent GT-AG introns are not always true GT-AG ones.

5. References

1) a) Stephen M. Mount, Nucleic Acids Res. 1982, 10, 459. b) Richard A. Padgett et al., Ann. __Rev. Biochem. 1986, 55, 119. c) Qiang Wu and Adrian R. Krainer, Mol. Cell. Biol, 1999, 19, 3225.
2) (U00921) I. Holzinge, et al., Immunogenetics, 1995, 42, 315.
(U18671) R. Yan, et al., Nucleic Acids Res., 1995, b23, 459.
(V00489) N. J. Proudfoot, et al., Cell, 1980, 21, 537.
(X52601) N. Kunze, et al.,Eur. J. Biochem., 1990, 194, 323.
(X73637) E. Barkhardt, Genomics, 1994, 20, 13.
(V80763) T. Hayashi, et al., Gene, 1997, 203, 231.
3) Peter D. Karp, Suzanne Paley and Jingchun Zhu, Bioinformatics, 2001, 17, 526.
4) E. Daniel, et al., J. Exp. Med. 1990, 171, 1.