

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

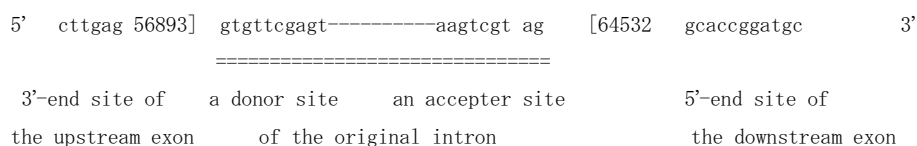
Intron is a portion of genomic DNA (and then RNA) sequence which is transcribed to an RNA but is to be finally eliminated by splicing. Many introns begin with the dinucleotide GT (or GU) and end with the dinucleotide AG, and therefore are called the GT-AG (or GU-AG) introns, which are described as those 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

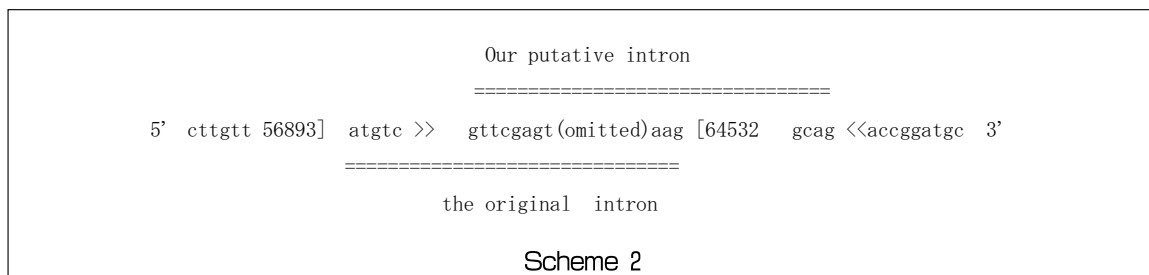
2.1 Notation

An intron is flanked by two exons, and have 5'-(or donor) and 3'-(or acceptor) 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 acceptor) (for example, Scheme 2)



2.2 Procedure

We collect sequences of introns whose both ends (5'-/donor and 3'-/acceptor) 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 interpret 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 acceptor 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]	cggccggggccggccctgcgg	L41919	atgaagcacgagccggccctgggta	[1597	gctatggcgac	
>>><<cccggactc 3459]	gctgtcggaccggcgacagcta	U07802	cggaccgggacagctaacaaagg	[3490	ctcc	
ggagctgaaa 11453]>>	gtgagtgaanaatggaggcga	U18671	tgfagctctctctcccaca(c)G	<<[11620	acggacacct	
c'tcaataga 14753]>>	gtgagatgaac'tgttcatt	U18671	ctftgtctgttgggtaaccag(g)	<<[16764	caggfggafga	
cc'tggagag>> g 133]	tgaggctcc	V00489	cag	<<gatgctctctcccaccaccagacctactcccg	[290	cactfcg
aacttcaag 454]>>	gtgagcggggccggggcg	V00489	actgacctctctc'gacag	<<ctc	[598	ctagccactg
ctccctgagct caggctgcag3428]>>	gtaagatgaaggagctgac	X17093	aggctcttctgtctaccoca(a)G	<<[3564	tcact'gacagt	
cgattcccag 1121]>>	gtacggcccggcc'gacctg	X52601	agtgctgtgtgataitctc	[1437	ctttctttccag <<at	
cagctgtg(a) 1470]>>	gtatcaacc'gctgccc'g	X53682	aatc'gctagccct'gctctacag	<<[2043	attccaaacca	
>>><<ggaggcacc 1330]	gtgagfggggacgggcaggac	X73637	cag			
	<<gtgagfgctacagtggtggagagacacatctgctctccatggcgtgggagcagtaatcagctgggacctggcctgcagccc		[2514	tgccc		
atccfttatg 410]>>	gtaactacannntcttca	X80763	gtaactacannntctt'cag	<<att	[436	atgctggcca
cggccgctcag(t) 415>>]	gtaggccggcgtctgcggga	X96924	cctctaccgccctccctccag<<[484	gtttggaaatg		
ccggatgcag 1764]>>	gtggtggggccgggagggga	X96924	gccccctgctgctgctgctccata	G <<[1846	ggccctgggggc	
>>><<	aagfgggag 2483]	jaagac'tgggatgatgaccaa	D42052	agactgggatgatgaccaaaatgat	[24859	tgaaaatggct
gtggfgctcc 5883]>>	t	D89870	<<[5885	gatgccccca		
gccgccggca 1268]>>	cc	M34677	<<[1271	ccagcccgggc		
ctaac'taag 29272]>>	c	U81031	<<[29274	tcgfggactg		
caaggaccag 705]	aaagctggaaagggctgtg	X75015	catggctgctcttttttttcag	[869	attgfggacct	
atcattcct 83]	nnnn	X80763	ntttttcag <<[t'g'c'ac			

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

aaagcaaa	530]	g>>	gfgggaccacagggggtgcga	AJ001563	cgctgtgccaacctgtctccctaca	[628	g	<<ggcagccccg
catgcca	82]	g>>	gtaagccaaccagggcctg	AJ001563	cgctcaccctccctctctctca	[201	g	<<cacctgagtt
aaagcaaa	330]	g>>	gfgggaccacagggggtgcga	AJ001564	cgctgtgccaacctgtctccctaca	[428	g	<<ggcagccccg
aaaaaag>>	g5395]	g>>	taagggctcaccctctgfgaa	AJ001981	actctctgctctctctccctaca	[5483	ctgggaaaaatg	
ccccacg>>	g5124]	g>>	taigtatgctctctctgfgctg	AJ001981	tccctctctgctctccctaca	[5233	cagtgfttatg	
gcccaggg>>	g5563]	g>>	taaatagctctctctgagccaa	AJ001981	attgctctctctctccctcag	[5649	tcctttacgac	
cttcaatt	6625]	g>>	gtaagtagagtcataaatgac	AL022069	aaatgagggtctctgtccccca	[8620	g	<<ccittcagac
agccccaca	57453]	g>>	gtgtgcccacagtggttcc	AL022069	ctgaactgtccccctgtcacctaca	[65669	g	<<actctctcgg aga
tcaca	16015]	g>>	gtttgtaaaagacg	AL022069	tggtctgtgfttgcattga	[18252	g	<<atttggcct
ccaagtc	156]	g>>	gtaagtagagggggccacc	L15346	cacctccctctctccctctca	[261	g	<<ggactaccctg
attccctca	386]	g>>	gtgtgtgaggtgtctctggg	L15346	cagacctctctctgaccactgtca	[476	g	<<ggatggagac
cagatgacg	2373]	g>>	gtgagggcggtctgccc	L19314	ccgctctctctctctggcccga	[3006	g	<<ctgctgctgag
atcgg>>	gta 1352]	g>>	agagtgatgatt	U90269	tgattgttctctctctctag	[1830	gataa	
tgagag>>	g 230]	g>>	tgaggctctctctctctctct	V00488	ccctcactctctctctcccccag	[348	atgttctctg	
gttccca	184]	g>>	gtaagtagaggaacactag	V00558	attgtttttatgttccaactca	[310	g	<<gfgcccagatg
tactca>>	g 5804]	g>>	taagtagcaggggagggcttc	X03562	ctctctgctctctctgagctccag	[7506	cagggccccca	
gagatcacc	612]	g>>	gtgagccccctgctctctg	X03900	gtctgatttctctctctccccc	[1336	g	<<gcaagcagca
aaatacac>>	g 1401]	g>>	tgagctctggcgcca	X03900	tgctctctctctctctctccccc	[1520	ctggccccggg	
ttcccgca	219]	g>>	gtgagctctggcctgcccggg	X63863	ggccctcactctctctctcccca	[335	g	<<accccaactg
actttaaa	1257]	g>>	gtaagacttccagatgtcaa	X64171	cacctctctctctctctctctct	[2308	g	<<cccagtgggc
cttttcca	727]	g>>	gtaaaactttcttaaaita	X67336	tctctctctctctctctctct	[1413	g	<<ctggacagac
agfgacg>>	g 463]	g>>	tgaggaggggtggtgggtgag	X70082	ggactctctctctctctctctct	[576	ctctggcctggg	
taaaaag>>	gt 571]	g>>	gattatggggggaactagagat	X70197	ccagggtgctctctctctctctct	<<gt	[675	gtccaggtgga
taaaaag>>	gt 250]	g>>	gattatggggggaactagagat	X70203	ccagggtgctctctctctctctct	<<gt	[354	gtccaggtgga
taaaaag>>	gt 250]	g>>	gattatggggggaactagagat	X70204	ccagggtgctctctctctctctct	<<gt	[354	gtccaggtgga
taaaaag>>	gt 250]	g>>	gattatggggggaactagagat	X70205	ccagggtgctctctctctctctct	<<gt	[354	gtccaggtgga
taaaaag>>	gt 250]	g>>	gattatggggggaactagagat	X70206	ccagggtgctctctctctctctct	<<gt	[354	atccaatgga
taaaaag>>	gt 571]	g>>	gattatggggggaactagagat	X70207	ccagggtgctctctctctctctct	<<gt	[675	gtccaggtgga
taaaaag>>	gt 339]	g>>	gattatggggggaactagagat	X70208	gaccatgctctctctctctctct	<<gt	[441	gtccaggtgga
catgcaa>>	g 2201]	g>>	taagtgacttctctctctctct	X75598	ctcattctctctctctctctctct	<<g	[6404	ctccgaagat
gcccagtg>>	gt 6506]	g>>	gagtgctctggtggtggatc	X75598	catactctctctctctctctct	<<gt	[7591	ctggggggggg
tcataatgg	10739]	g>>	gfgagtttcccaacagttgc	X91233	attaatctctctctctctctct	[13493	g	<<aatgtaacag
tgfgtga	26818]	g>>	gtaagcaagcgtg	Y09912	atacttctctctctctctctct	[28883	g	<<gagaagcgt
ttaaaaa	17406]	g>>	gtagataatccccccaa	Y09912	taaggcaattttctctctctct	[24530	g	<<tcccagttcc
cttcaaca	13503]	g>>	gtgagccactgcttctgca	Z83307	tgctatttctctctctctctct	[14194	g	<<gactcatttc

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

gctggcagg>>	g	97022]	taagcagtcacctaacaacagg	AL022069	ccacaattgctcaattatacag	<<a	[98562	fttaccctftt
catccicaa>>	g	26233]	tgagtatgtgaggagccacc	AL022069	ctaaataggatctgtttttcag	<<a	[27983	gccaagctggg
atggaccaaa11094]	tg>>		gtaatttcaataat	D42052	ttcttttttaactgtaactttg		[11839	ag <<gaatgaaca
cc1gggcag>>	g	265]	tfgtataaggtataagag	V00505	ctgggctgttttctaccctcag	<<a	[394	ttaactgggggt
ct1gggcag>>	g	2171]	taagcattggtctcaatgca	V00508	atcigtatggtgtcatttcag	<<a	[2294	ctcctcgtftg
cc1gacgg>>	gt	421]	gagtccctgcgcgcgggg	X63863	tgaggggaggtcccccaacag	<<gc	[577	gacctggagag
tttccgg>>	gta	654]	agctacaggtctgtggcca	X63863	atcctcgttcgggtctcag	<<gtt	[748	aaocg
cc1cagaag>>	g	24751]	taacccccaccagaaaacaa	Y09912	gaatgtaattctgcaatttcag	<<a	[26702	gccaatcgaa

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	gacggatcag	SW:Q13705
gcaaggctgc39864]	lagacg>>	gtaagtagggatggcag	AB008681	ctgccaggctctctctcctag	<<[39963	gaccctggat	SW:Q13705	SW:Q13705
aaaatc>>	gtaa24575]	gtaataggtatgtgatt	AC004237	ttaaaggattcgtttattttaa	[28194	ag <<gttctgtg	SW:O15066	SW:O15066
tagag>>	gtaag19057]	gataatgaaatggggagt	AC004237	taattttttttt	<<gtaaaagaag	[24477	acctgatggg	SW:O15066
gtactfgaag10966]	gctacaatg>>	gtaagfgaatct	AC004237	aactgattctctcatctttcag	<<[18682	ggactatttt	SW:O15066	SW:O15066
gggaaggacg1240]>>	gtgagtgtccacgccccttc	AF020276	ctcctcgtgtgtgtatctct	[1472	ag <<gacagaatt	PRF:2401341A	PRF:2401341A	PRF:2401341A
caeggctgc2579]>>	gtaagtgtatggcggcggcgggt	AJ001977	ggctttttttgtttaccaccag	<<g	[2687	cagcaacagtg	PRF:2314363	PRF:2314363
ctftgtaaag>>	g	2734]	tgagattctggggagc1gaa	AJ001977	tgttcagactattggtctgtag	<<[2898	cc1gagacagc	PRF:2314363
agcaccacac132]>>	gtgagtgccctcgggagaggag	AJ006854	gctgacctaccgctctctctca	[266	g <<gaaaagc1ga	SW:P19440	PRF:2413277A	SW:P19440
ggggatcag12735]	gtaagtcacagttcaacctgc	AJ224869	atgcttgc1gtaattgggaagtgaatg	[4844	tcattcctttgcctcttttgcag	<<atata	PRF:2413277A	PRF:2413277A
ccggctgg>>	gt28072]	laagattccccgggacaccca	AL022069	gatgatctttgtgtttgaca	[46997	g <<gtgctggcat	SW:Q15349	SW:Q15349
gaagcctgag12395]	jaa>>	gtaagtga gaa aaaaactag	AL022069	catctgtttctctccctftttt	[15977	ag <<catctctct	SW:Q15349	SW:Q15349
cg1gctcgg21888]>>	gtgcggagggcggcggcggc	D26607	tcgatccactgctctttccga	[21975	cag <<gatcagca	SW:P29474	PIR:JN0684	SW:P29474
gaaagaa>>	gtg2456]	agtcggggcagcgcctcccc	D43639	aatcggctgctgtgttttccag	<<[2604	gtggataaagt	PIR:JN0684	PIR:JN0684
tc1ggccccg1337]	gatgggtacggcgaacccagaccggcctccccgca>>	gt	D50739	gagcctg1gctgtttgctttacag	<<[2329	tc1cccgcttc	SW:P43699	SW:P43699

(Table 4)

gaggcagagg4662]acctgcagg>> cagctacag>> tcaggaa>> cga349] acttccgaccgtctctgftaacgcggatggcgag>> atgaaagag>> ccg 614]>> aacac 1477] acggagatg>> gctaaggtgg1963] >> tgcactca>> gaggagaca>> accacag>> gta1386] agttgggtggcttgagctg agctttgg>> ctfcaaggtga 9234] gtagcaaaaggacacaaaggaa gg>> gg>> atgaccactgaggattacaagaagct>> gaggccagcg 720]>> agaagacaa>> caggctgcgt3041]>> ccaacttacct438]>> caccctgag>> gccaactfacc437] t>> aggcgtcaag807] a>> g29026] tga gttattagactctgggc gaaagaag81]>> agcggagcccg270]>> ctcg>> ctgtgtgact4627] cag>>	gtagagccaaccg taagggcttccctagtcctaa agtatggggccctgggggtctggg ftaacgcggatggcgag>> g 2836] tga g c g c t g c a c t c c c t c c c t U02948 catgctgtctcttgcctgccccca [1319 U02948 ggggagaccaggggctccatgccc U19759 cgttccatggttcatctgacttcagg [2817 U24578 ccacttttggcagctcatccccgc [16835 U24578 tgttgcactgtcccttgnccccctgc [17111 U25441 ccttcccccttctcatccacag <<ggg [907 gtagagccaaccg [363 M24543 atctctcgtctgcttctccccag <<[3351 M94363 tgcgccggtctctgtccccgcacag <<[508 U00921 catcag <<[3593 cogtgtcag <<ga [1823 gggagagctgctccatgccc [1823 tag <<agctacag [2817 ag <<ccaaggatg [16835 agnccccag <<ta [17111 <<ggg [907 gtagagccaaccg [363 M24543 atctctcgtctgcttctccccag <<[3351 M94363 tgcgccggtctctgtccccgcacag <<[508 U00921 catcag <<[3593 cogtgtcag <<ga [1823 gggagagctgctccatgccc [1823 tag <<agctacag [2817 ag <<ccaaggatg [16835 agnccccag <<ta [17111 <<ggg [907	gtagagccaaccg taagggcttccctagtcctaa agtatggggccctgggggtctggg ftaacgcggatggcgag>> g 2836] tga g c g c t g c a c t c c c t c c c t U02948 catgctgtctcttgcctgccccca [1319 U02948 ggggagaccaggggctccatgccc [1823 U19759 cgttccatggttcatctgacttcagg [2817 U24578 ccacttttggcagctcatccccgc [16835 U24578 tgttgcactgtcccttgnccccctgc [17111 U25441 ccttcccccttctcatccacag <<ggg [907 gtagagccaaccg [363 M24543 atctctcgtctgcttctccccag <<[3351 M94363 tgcgccggtctctgtccccgcacag <<[508 U00921 catcag <<[3593 cogtgtcag <<ga [1823 gggagagctgctccatgccc [1823 tag <<agctacag [2817 ag <<ccaaggatg [16835 agnccccag <<ta [17111 <<ggg [907	SW:P01308 PIR:S49530 SW:P07288 PIR:45023 PRF:2402239A SW:P11686 SW:P11686 SW:Q12986 PRF:1006211A PRF:1006211A PRF:1705199A PRF:1705199A SW:901524 PIR:A43359 SW:P52333 SW:Q99490 SW:O15231 SW:P05217 PRF:2401351A SW:P10317 SW:P08686 SW:P08686 SW:P08686 SW:P52656 PRF:2218269A PRF:1505244A PIR:I37515 PIR:G02334 SW:P48067
--	---	---	---

(Table 4)

tggattttatg5264]>>gtaaatagtaactataaaatt Z96810 ctggaagaataatattag <<caagcact [5521 caatgtactct SW:P48067
 tggg2237]>> gtaag Z97370 cccag <<agccatctcccagccccaccgcccccatctgctggcctggttcctctf [2400 gagcctgt PRF:1311286A
 ctftcag>> gta 1353] aggagfttaacattgaaatg X07963 ctfttaattatagtcctcag <<igt [1483 gacatccagat

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

A capital letter is a nucleotide to be inserted.

ftgaaaaatt 1440]>> gtaagtttt AB010084 acctaattaacaaaatt CTAATCATTCTTATAG <<[1569 cttttactaaa SW:Q31241
 catcccaag3291]>> gtaagtaaa AF026029 agagatacaatgacaattctttca G <<[3772 ggttgcgtat PRF:2201493A
 aggGttacag>> g 9909] tgggagtgccctgttagtcc U24578 cctccaccctctcccctcctcagtag <<[10045 cctmagatcca SW:01028
 ctgctacaag 658]>> G taagcactgctg U41163 gccctccaccctccag <<[749 gacgccatcat PIR:G02277
 cgacaaccag1303]>> Gtttgcattgggctctgggacag U41163 tcggcctgagctagcctggccacag <<[1414ttttagtgtgt PIR:G02277
 atctcttcag9452]>> gtgcccctgggacgggtgg U70065 cactcccctgccccaccacacga G <<[10839actatgagctc SW:P52333
 ggacagcatt15063] cGcgag>> gtaagagaaggctcag U81031 ctcttactgcccctctcccgttaca [21547 gag <<gctgtgat SW:Q99490

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

ctccacag9615]>> gtctagagccagcaggaac U18671 ttcttccacacctctctctctcg <<[9716 agccfttftggs PRF:2108377A
 ccccgctgcg4187]>> gtagcggcggggacca U28054 cccgacctatctcggctccatctacg <<[4389 ggctccgagggca SW:P26927
 acaagccgca5006]>> gtgagctccctggtgctccgg U28054 cgaftttgctcccgcctccgctcg <<[5134 gttcacgfttia SW:P26927
 ataccgfggg 3309] gg>> gtaagaaagggccccctgacag U48705 tctcttctctggtcccctctcttc [3841 cg <<gactgacagt SW:Q08345

(Table 6)

aagctaacag>> gf68]	gtatctatctgtccctgggct	L24038	ccctgtagtgcttgcaccgcccggg	<<[9797	acatctccta	SW:P10398
ctctcag>> gf63]	gtgggatcaagggttacaaga	L48213	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
ctctcag>> gf63]	gtgggatcaagggttacaaga	L48214	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
ctctcag>> gf63]	gtgggatcaagggttacaaga	L48215	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48216	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48217	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48220	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48221	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48931	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
	gtgggatcaagggttacaaga	L48932	ggcactgactctctcggccattggg	<<[420	tctatctccc	PRF:070522A
aagctaacag>> gf63]	gtatctatctgtccctgggct	U01337	ccctgtagtgcttgcaccgcccggg	<<[9797	acatctccta	SW:P10398
agcctgag>> gf68]	aagfgaatgcttgagcccaggg	AF039597	tcttcttctctctctccctag	[186	ctctg	<<<gactgt	
ctctcag>> gf63]	gagcccacacacctctccc	U42588	ctcggcattggtcacaatgagac	[414	aactg	<<<gttcga	
ctctcag>> gf63]	gagcccacacacctctccc	U42589	ccctggcattggtcacaatgagac	[414	aactg	<<<gttcga	
ctctcag>> gf63]	gagcccacacacctctccc	U42591	ccctggcattggtcacaatgagac	[414	aactg	<<<gttcga	

Table 7. GC-AG introns (20)

acaagtggaag14818]>>	gcaagtaaatgaaat	AF001295	ccctaccatgfgtctag	<<[16913	gftacagtgac	SW:P17643
ccacaagaag4024]>>	gcaagtagaggga	AF019084	tcctcccccataatcag	<<[4342	tatgaggagct	SW:P35908
tcfgaagaag2522]>>	gcaagtgacaca	AJ000263	actccccacccttccag	<<[3282	gatgfggactg	SW:P35908
agatgaactg620]>>	gcatgfgctgagtc	AJ000673	cttttacttgaagcag	<<[2415	gattttatgag	SW:Q13241
agtttcagag13787]>>	gcaagtgctcattt	D42052	aatgtgtttatgtttttag	<<[14272	gcatagagaag	SW:P50900
gggaccccaag30192]>>	gcaagttctgcc	L23982	atctgtgftttctcacag	<<[30441	ggtaaccctgg	SW:Q02388
tftggatcag2163]>>	gcaacctgccctccc	L46590	atcccccttctccacag	<<[2259	taatggggggcc	SW:A54183
cagaagaagaag3315]>>	gcgagtggggggtggagaggggg	U28054	tfgatctc-tgctcttctgctcccccag	<<[3393	actacatacagg	SW:P26927
cagaagaagaag5243]>>	gcaagtggggggtggagaggggg	U37055	tfgatctc-tgctcttctgctcccccag	<<[5321	actacgtacagg	SW:P26927
gaaacacccag1017]>>	gcaagtgctctcttctttagct	U51038	gtttcgtatgctcattttaaacactacag	<<[2428	agtggtatccct	PIR:A47297
catgatgaag118086]>>	gcaagtggtactcagccca	U82828	tfttggctcttttttaattggtat	<<[119109	agagacgggaat	PIR:A4310
ctgcccccaag44163]>>	gcaagtgaccacacaaatctgcc	U89335	tcctccccctccctccacag	<<[45334	ggttttgaaagg	SW:P31695
tgctactcag2914]>>	gcatgfgccccacccttcccaca	X56494	atgctctgctccctcccccag	<<[3162	atgctgggagag	SW:14618
gggttcggag2679]>>	gcaagtgccccgttcttcccctgg	X56494	cccccaacttctgctccatcag	<<[2764	gtttgatgaaa	SW:14618
ttctgttaag2112]>>	gcaagttaccagatgattccta	X94208	ttttcttggatttcttttag	<<[553	tatcgagaaaa	SW:P12270
tacatcgcaag937]>>	gcgagtgccccagtgcccgcat	Y00486	ctgtcaacttaccctgacag	<<[1931	gacctagactcc	SW:P07741
tcfgaagaag2545]>>	gcaagtgacacacagggattgaga	Y13621	acactccccacccttccag	<<[3303	gatgfggactg	SW:P35908
caaggcccaag22664]>>	gcaggctctgctcggccctccc	Z20656	cc-tgcttctgcttccctgaaag	<<[23001	gcaaacctggga	SW:P13533
tccaggcaat2098]>>	gctggagtaactggcgggttgca	AF026029	aaattatttttcttccigata	G <<[2808	gctggccccgggt	PRF:2201493A

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
2424gtgagttgctacagtggtggagagacgacatctgctccatgggctggtggccgacagtaatctcacgctgggacctggcctgcagcccc2513
[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 >>gtt-----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 >>gtt---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 it⁴⁾ 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 notes³⁾.

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. Holzinger, 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.