Detection of Intron, Exon, and Intergenic DNA in Human Genome on the Basis of Quantification Method II

Hiroki Fukasawa[1] Shigehiko Kanaya[1] Yoshihiro Kudo[1]

m96041@eie.yz.yamagata-u.ac.jp kanaya@eie.yz.yamagata-u.ac.jp ykudo@eie.yz.yamagata-u.ac.jp

[1] Department of Electric and Information Engineering, Faculty of Engineering, Yamagata University, Yonezawa, Yamagata-ken 992, Japan

The development of methods to detect genes in DNA sequences is important for genome analysis. In the previous study, we have developed measures which reflect the species-specific diversity of codon usage among genes in prokaryotes. Using the measure developed, we could also success to predict protein-coding regions in *Escherichia coli* genome. In the present study, we examined the procedure for detection of intron, exon, and intergenic DNA in Human genome by Quantification Method II.

To discirminate boundaries (1) between 5'-intergenic DNA and 3'-exon, (2) between 5'-exon and 3'-intron, (3) between 5'-intron and 3'-exon, and (4) between 5'-exon and 3'-intergenic DNA, we constructed data sets consisting of DNA sequences including these boundaries (Groups 1 to 4, respectively) and intra DNA sequences of intron, exon, and intergenic DNA (Group 5). The number of sequences is 774 for Group 1, 2327 for Group 2, 2327 for Group 3 , 779 for Group 4, and 1551 for Group 5. The gth discrimination function was represented by *Eq.* (1).

$$F(g) = \sum_{i=1}^{2n} \sum_{j=1}^{4} a_{ij}^{g} x_{ij} + c^{g} \tag{1}$$

Here, x_{ij} represents dummy variable for the jth nucleotide (j=1 to 4 correspond to A, T, G, and C, respecitvely) at the ith position. For example, if nucleotide at the ith position for a sequence is T, then $x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$, and x_{i4} =0. The weight for the gth discrimination function is denoted by a_{ij}^{g}. The discrimination functions were developed for pairwise between a group and the remaining groups according to the inter-group distance among the five groups as follows. In the first step, Group 2 [F(1)≥0] was discriminated from the other groups [F(1)<0] by F(1), in the second step, Group 1 [F(2)≥0] was discriminated from the remaining groups (Groups 3 to 5) by F(2), in the third step, Group 4 [F(3)≥0] was discriminated from the Groups 3 and 5 by F(3), and in the final step, Group 3 [F(4)≥0] was discriminated from Group 5 by F(4).

Table 1 shows the weight for the four discrimination functions for 16-nucleotide sequence with boundary at the center. Discrimination rates for these four functions are 0.949 for F(1), 0.922 for F(2), 0.869 for F(3), and 0.751 for F(4). The nucleotides at the 5th position as well as 3 nucleotides around the boundary play an important role in the discrimination between 5'-exon and 3'-intron (see Range of F(1) in Table 1).

Table 1: Parameters for discrimination functions [F(1), F(2), F(3), F(4)]

			F(1)	——	F(2)	——	F(3)	——-	F(4)	——
Pos.	Nt.	Freq.	Weight	Range	Weight	Range	Weight	Range	Weight	Range
-8	A	1488	-0.0003	0.0101	0.0336	0.0673	0.0835	0.1372	-0.0016	0.0689
	T	2004	-0.0036		-0.0337		-0.0537		-0.0073	
	G	1708	0.0065		0.0336		0.0403		-0.0373	
	C	2558	-0.0013		-0.0156		-0.0334		0.0316	
-7	A	1491	0.0248	0.0591	0.0532	0.0867	0.0150	0.0890	-0.0352	0.0767
	T	2102	-0.0181		-0.0335		-0.0337		-0.0090	
	G	1571	0.0371		0.0321		0.0552		-0.0230	
	C	2594	-0.0220		-0.0229		-0.0147		0.0414	
-6	A	1484	0.0263	0.0453	0.0197	0.1281	0.0800	0.1282	-0.0560	0.0991
	T	2051	-0.0074		-0.0430		-0.0432		-0.0238	
	G	1731	0.0135		0.0851		0.0520		0.0431	
	C	2492	-0.0190		-0.0355		-0.0482		0.0230	
-5	A	1640	0.0112	0.0391	-0.0013	0.0774	0.0935	0.1481	-0.0054	0.0864
	T	2320	-0.0156		-0.0261		-0.0546		0.0131	
	G	1409	0.0235		0.0513		0.0382		-0.0601	
	C	2389	-0.0065		-0.0040		-0.0337		0.0263	
-4	A	1751	0.0123	0.0233	-0.0282	0.0535	-0.0223	0.0703	-0.0121	0.0863
	T	1486	-0.0045		-0.0137		-0.0060		-0.0410	
	G	1966	-0.0111		0.0027		0.0479		-0.0171	
	C	2555	0.0028		0.0253		-0.0181		0.0453	
-3	A	1707	0.1428	0.3484	0.1933	0.3514	-0.0655	0.3895	-0.0743	0.3534
	T	1940	-0.2056		-0.0877		0.2675		0.2110	
	G	1115	0.0922		0.2172		-0.0371		-0.1424	
	C	2996	0.0175		-0.1342		-0.1221		-0.0413	
-2	A	4560	-0.0241	0.1351	-0.0960	0.3257	-0.0290	0.2320	0.0815	0.3148
	T	822	0.0371		0.2296		0.0319		-0.2334	
	G	1298	-0.0279		0.0459		0.1498		-0.0447	
	C	1078	0.1072		0.1757		-0.0822		-0.1127	
-1	A	1313	-0.2384	0.3300	0.1569	0.5255	0.3559	0.4487	0.1479	0.4561
	T	630	-0.0996		0.2054		-0.0040		-0.3082	
	G	4917	0.0917		-0.1388		-0.0776		0.0113	
	C	898	-0.0837		0.3867		-0.0928		-0.0618	
1	A	1915	-0.1729	0.3986	0.2873	0.4165	-0.1799	0.2498	0.2009	0.3942
	T	752	-0.1395		0.0071		0.0285		-0.1688	
	G	4224	0.1535		-0.1292		0.0699		-0.0213	
	C	867	-0.2451		-0.0112		0.0319		-0.1933	
2	A	1111	-0.2512	0.4993	-0.1012	0.2084	-0.0060	0.0231	-0.1424	0.2371
	T	4281	0.2154		0.0841		0.0075		0.0947	
	G	1244	-0.2839		-0.1243		-0.0156		-0.1014	
	C	1122	-0.2583		-0.0829		-0.0055		-0.1080	
3	A	2166	0.1272	0.2714	-0.0613	0.1862	0.0539	0.0862	-0.0361	0.1643
	T	1281	-0.0783		-0.0705		-0.0069		-0.0838	
	G	3043	0.0025		0.1065		-0.0322		0.0806	
	C	1268	-0.1442		-0.0797		-0.0079		-0.0470	
4	A	2782	0.1263	0.2437	-0.0320	0.0900	0.0294	0.0752	-0.0328	0.0847
	T	1385	-0.1174		0.0018		-0.0298		-0.0290	
	G	1934	-0.0567		0.0580		-0.0459		0.0519	
	C	1657	-0.0477		-0.0156		0.0292		0.0187	
5	A	1496	-0.1080	0.2708	0.0319	0.1224	-0.0708	0.1807	-0.0003	0.0783
	T	1347	-0.1054		0.0106		-0.0385		-0.0343	
	G	3186	0.1570		-0.0557		0.0957		-0.0092	
	C	1729	-0.1138		0.0667		-0.0850		0.0440	
6	A	1442	-0.0490	0.1228	0.0113	0.0877	-0.0143	0.0346	-0.0687	0.1042
	T	2390	0.0738		-0.0445		0.0082		-0.0121	
	G	2024	-0.0252		0.0432		-0.0165		0.0299	
	C	1902	-0.0287		0.0015		0.0181		0.0355	
7	A	1849	-0.0002	0.0296	-0.0213	0.0353	-0.0267	0.0431	-0.0039	0.0465
	T	1538	-0.0154		0.0140		0.0116		-0.0306	
	G	2363	0.0141		-0.0016		0.0164		0.0095	
	C	2008	-0.0047		0.0108		-0.0036		0.0159	
8	A	1739	-0.0278	0.0511	-0.0055	0.0274	-0.0095	0.0137	0.0011	0.0104
	T	1885	-0.0135		0.0174		0.0014		-0.0068	
	G	1933	0.0116		-0.0101		0.0042		0.0036	
	C	2201	0.0233		-0.0016		0.0026		0.0018	
c^g			-0.3056		-0.5445		-0.2212		0.0921	