PFcDNA: Prediction of a Full Length Gene from Partial Sequence

SooYoung Cho[1] singylu@ihanyang.ac.kr
Myungguen Chung[1,3] aobo@ihanyang.ac.kr
Seung-baek Han[2] jhnvalor@korea.ac.kr
Hyun Kim[2] Kimhyun@korea.ac.kr
YoungSeek Lee[1] yslee@hanyang.ac.kr

[1] Dept of Biochemistry, Han-yang University, Seoul, South Korea
[2] Dept of Anatomy, Medical school, Korea University, Seoul, South Korea
[3] Bioinformatics research team, Electronics and Telecommunication Research Institute, Dae-jun, South Korea

Keywords: gene prediction, cDNA, EST

1 Introduction

We obtained 2500 EST sequences from the cDNA library which was constructed from developmental stage of rat brain tissue. We could not find ant homologous sequence for most of the EST sequences from 'nr' database by using BLAST. BLAST was not sufficient to find full length sequence from an EST sequence because relatively small number of full length rat genes in the public databases. We change longer query sequence for BLAST than before using PFcDNA. It was possible to predict several completed full length genes from EST sequence and cloned full length genes by using a predicted sequence. In order to predict obtained gene function, we extract the annotation and function modified gene-ontology (GO) of the gene.

To find out the full length genes and function prediction for a short EST sequence.

1) Each of sequence was BLAT to Rat whole genome, and then, 100Kb region of include matched sequence was extracted from Rat whole genome.
2) Genes are predicted by using GENSCAN from 100Kb genome sequence and the predicted genes are BLAST against several full length cDNA library databases.

2 Implementation

PFcDNA consists of four main modules, i.e. Sequence Pre-processing module, Upstream Prediction module, Upstream Evaluation module, annotation module. The main purpose of 'Sequence Pre-processing module' is data readjustment. We have to trim out error or a supplement sequence in the experiment procedure. 'Upstream prediction module' is genomic alignments which take the goal at finding relative genomic position for partial sequence. For the removal purpose of false positive prediction, we construct 'Upstream Evaluation module'. 'Annotation' confirms function for predicted sequence.

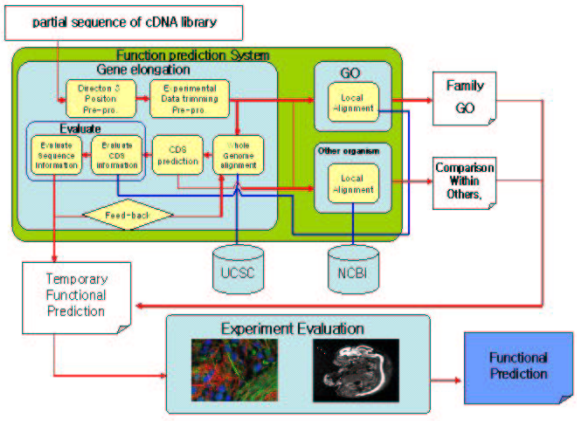

Figure 1: Outline schema of PFcDNA.

All four module are independent module using Perl(`http://www.perl.com/`) and MySQL (`http://www.mysql.com/`).

3 Results

We got several partial cDNA sequences from the cDNA library which was constructed from developmental stage of rat brain tissue using RT-PCR (Reverse Transcript Polymerase Chain Reaction). The BLAST searching for each of the sequence is not available to find gene function. We execute 'PFcDNA' application for finding full cDNA sequence which is aligned to whole genome sequence (UCSC, June 2003) [3] by using BLAT. Then we annotated gene function by Modified-GeneOntology(MGO) [1] which is constructed for finding function from results of the BLAST(Table 1).

Table 1: Prediction of gene function using MGO from BLAST result.

	gene class																
gene id	behavior	apoptosis regulator activity	binding activity	chaperone activity	defense/immunity protein activity	enzyme activity	cellular process	signal transducer activity	structural molecule activity	transcription regulator activity	transporter activity	development	obsolete	physiological processes	cell	extracellular	sub total
A0075a			75			69	58	1	22	12	5	13	1	210	190	8	664
A0187			142	23		91	98	10	43	13	5	21	1	396	384	8	1235
A0222b			139	23		89	98	8	43	13	5	21	1	390	381	8	1219
A0545b			142	23		91	98	10	43	13	5	21	1	396	384	8	1235
A1258a-24E16-2-1			133	23		89	93	7	38	12	5	18	1	351	334	8	1112
A1270b-24P3-4-2			8			4	8				2			6	2		30
A1294-25E14-1			139	23		89	98	8	43	13	5	21	1	372	381	8	1201
A2236a-059E14-2-1			169	23		91	98	10	43	33	5	23	2	437	411	8	1353
A2266a			109	22		69	70	1	22	12	5	17	1	250	206	8	792
B0200			82			69	70	1	22	12	5	17	1	222	198	8	707
B0479b			109	22		69	70	1	22	12	5	17	1	250	206	8	792
B0484			142	23		89	98	10	43	13	5	21	1	394	384	8	1231
B0691b			142	23		91	98	10	43	13	5	21	1	396	384	8	1235
B0692			26			24	32	1	9	1	5	11	1	57	87	1	255
C2255b	6	2	266	23	3	165	229	61	81	36	12	46	3	687	526	19	2165
C2265			75			69	58	1	22	12	5	13	1	210	190	8	664
C2404d			142	23		89	98	10	43	13	5	21	1	394	384	8	1231
C2417b			133	23		89	93	7	38	12	5	18	1	351	334	8	1112
C2443d			11			10	14				2			6	2		45
C2444e			141	23		89	98	10	43	13	5	21	1	392	381	8	1225
C2467b			141	23		89	98	10	43	13	5	21	1	392	381	8	1225
C2467d			73			69	52	1	10	11	5	11	1	195	117	8	553
C2492d			133	23		89	93	7	38	12	5	18	1	351	334	8	1112
C3328a			141	23		89	98	10	43	13	5	21	1	392	381	8	1225
C3338a			43			69	36	1	10	3	5	11	1	171	103	8	461
D0627a	4	1	241	23		107	209	40	81	35	7	46	3	613	498	8	1916
D1401a	6	2	266	23	3	165	236	61	91	36	12	46	3	700	563	19	2232
D2594a			139	23		89	98	8	43	13	5	21	1	390	381	8	1219
D2594b			133	23		89	93	7	38	12	5	18	1	351	334	8	1112
D2595a			142	23		89	98	10	43	13	5	21	1	394	384	8	1231
sub total	16	5	3777	504	6	2479	2788	322	1103	419	160	595	35	10116	9225	239	31789

4 Discussion

We take 42 full length genes for finding gene function constructed from developmental stage of rat brain tissue using RFcDNA. We want to perform the experiment for evaluation results of RFcDNA and make an in-Situ hybridization experiment that is confirms locations of expressed gene in a living thing [2].

References

[1] Chung, M., Prediction of a full length gene from partial sequence, In *Dept. of Biochemistry 2003, Hanyang University: Seoul*, 86, 2003.

[2] Gilbert, S.F., *Developmental Biology, 7th ed.*, Sinauer Associates, 838, 2003.

[3] UCSC `http://genome.ucsc.edu/`