

GIGGene installation instructions

Alexandre Tchourbanov

*Department of Computer Science, College of Information Science and Technology,
University of Nebraska at Omaha, Omaha, NE 68182-0116, achurbanov@mail.unomaha.edu*

March 3, 2004

1 Basic use

1.1 Installation instructions

1. Download the latest version of the Java Software Development Kit (SDK) from <http://www.sun.com> and install.
2. For BLASTN stand-alone sequence comparison software installation visit <ftp://ftp.ncbi.nih.gov/blast/>. Instructions to run BLASTN program are available at <http://www.ncbi.nlm.nih.gov/Education/blasttutorial.html>.
3. Download the GIGGene program archive `GIGGene.1.0.zip` from <http://bioinformatics.ist.unomaha.edu/~achurban/> and unzip it in its home directory. For example, suppose we use home directory `C:\GIGGene` under Windows while on Unix we go with `/home/GIGGene`.
4. Set the environment variable `CLASSPATH` point to the GIGGene home directory, plus the BioJava and PostgreSQL JDBC driver JAR packages and the current working directory (`.`). In order to do this:
 - For UNIX system type `export CLASSPATH=CLASSPATH:./home/GIGGene/biojava-1.30-jdk14.jar:/home/GIGGene:/home/GIGGene/postgresql.jar`
 - To setup `CLASSPATH` variable on the Windows XP machine, go to `Start → Settings → Control Panel → System → Advanced → Environment Variables` and then choose `New`, for `Variable Name:` input `CLASSPATH` and for `Variable Value:` enter `.;C:\GIGGene;C:\GIGGene\biojava-1.30-jdk14.jar;C:\GIGGene\postgresql.jar`
5. General command structure to start GIGGene program is:

```
java BlastObject.BlastObjectParser <path/Param> <path/Blast> <path/Output>
```

where `path/Param` is path to file with parameters, `path/Blast` is path to file with BLASTN search results and `path/Output` is the output file to be written in FASTA format (<http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml>).

For example, in order to start the application for the test case under Unix control, type the following command from the GIGGene directory:

```
java BlastObject.BlastObjectParser  
/home/GIGGene/BlastObject/Parameters/clusterResultsParse.xml  
/home/GIGGene/BlastObject/NM_003890.1.blast  
/home/GIGGene/test_out.fa
```

The result of the command on the screen should look, as shown in subsection 1.3. Also, the program writes exons and introns predicted into FASTA file `/home/GIGGene/test_out.fa`, with description elements explained in subsection 2.3.

1.2 Input BLASTN file

The input file for the GIG0gene program is the result of BLASTN search, represented as a text file. For the purposes of the BLASTN runs we have compiled DNAA11.fa database (<http://bioinformatics.ist.unomaha.edu/~achurban/>) of finished human genomic clones extracted from PRI files at NCBI ftp://ftp.ncbi.nih.gov/genbank/. DNA databases for other organisms could also be compiled.

We use human mRNA Reference Sequences to search DNA database. Example of BLASTN search result for the test run, shrunk in the middle, is listed below:

BLASTN 2.2.3 [Apr-24-2002]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= gi|4503680|ref|NM_003890.1| Homo sapiens Fc fragment of IgG binding protein (FCGBP), mRNA
(16,382 letters)

Database: DNAA11.fa
155,760 sequences; 3,822,543,938 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value
gi 5080755 gb AC007842.1 AC007842 Homo sapiens chromosome 19, BA...	2476	0.0
gi 27733931 gb AC138626.1 AC138626 Homo sapiens chromosome 19 cl...	2476	0.0
gi 21617636 gb AC011536.6 AC011536 Homo sapiens chromosome 19 cl...	1231	0.0
gi 4321125 gb AC006950.1 AC006950 Homo sapiens chromosome 19, co...	1136	0.0
gi 24137484 gb AC006548.22 AC006548 Homo sapiens chromosome 22 c...	48	0.21
gi 15145527 gb AC091650.4 AC091650 Homo sapiens BAC clone RP13-4...	46	0.84
gi 11386322 gb AC025594.5 AC025594 Homo sapiens chromosome 7 clo...	46	0.84
gi 11386320 gb AC024952.4 AC024952 Homo sapiens chromosome 7 clo...	46	0.84

>gi|5080755|gb|AC007842.1|AC007842 Homo sapiens chromosome 19, BAC 331191
(CIT-B-471f3), complete sequence.
Length = 174009

Score = 2476 bits (1249), Expect = 0.0
Identities = 1249/1249 (100%)
Strand = Plus / Minus

```
Query: 63      ggattgaccaggaggcttcagtggacctcaagaacactggcagagaggaattcctcaca 122
              |||
Sbjct: 46549  ggattgaccaggaggcttcagtggacctcaagaacactggcagagaggaattcctcaca 46490
```

Query: 123 gccttcctgcagaactatcagctggcctacagcaaggcctacccccgcctccttatctcc 182
 |||
 Sbjct: 46489 gccttcctgcagaactatcagctggcctacagcaaggcctacccccgcctccttatctcc 46430

.....

effective length of query: 16359
 effective length of database: 3,818,961,458
 effective search space: 62474390491422
 effective search space used: 62474390491422
 T: 0
 A: 40
 X1: 6 (11.9 bits)
 X2: 15 (29.7 bits)
 S1: 12 (24.3 bits)
 S2: 22 (44.1 bits)

1.3 The test run result

The result of the test run should look as following:

GIG0gene-version 1.1
 ## date: Wed Mar 3 11:37:24 2004

gb AC007842.1 gi 5080755	GIG0gene	exon	52806	52868	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	46549	52805	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	45301	46548	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	42975	45300	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	42625	42974	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	36885	42624	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	36314	36884	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	34032	36313	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	33427	34031	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	32500	33426	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	31932	32499	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	24566	31931	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	23973	24565	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	21184	23972	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	20589	21183	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	20471	20588	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	20271	20470	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	18396	20270	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	18254	18395	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	14807	18253	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	14469	14806	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	13159	14468	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	12585	13158	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	12191	12584	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	11574	12190	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	10846	11573	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	10302	10845	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	8732	10301	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	8139	8731	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	5246	8138	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	exon	4651	5245	.	-	.
gb AC007842.1 gi 5080755	GIG0gene	intron	4533	4650	.	-	.

gb AC007842.1 gi 5080755	GIGOgene	exon	4333	4532	.	-	.
gb AC007842.1 gi 5080755	GIGOgene	intron	2129	4332	.	-	.
gb AC007842.1 gi 5080755	GIGOgene	exon	1987	2128	.	-	.

GIGOgene-version 1.1
date: Wed Mar 3 11:37:58 2004

gb AC011536.6 gi 21617636	GIGOgene	exon	37532	38124	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	34639	37531	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	34044	34638	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	33926	34043	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	33726	33925	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	31499	33725	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	31357	31498	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	27945	31356	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	27607	27944	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	26297	27606	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	25723	26296	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	25329	25722	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	24712	25328	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	23988	24711	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	23444	23987	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	21873	23443	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	21280	21872	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	18387	21279	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	17792	18386	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	17674	17791	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	17474	17673	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	15248	17473	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	15106	15247	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	11694	15105	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	11356	11693	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	10038	11355	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	9464	10037	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	9076	9463	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	8459	9075	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	7733	8458	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	7189	7732	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	5597	7188	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	5025	5596	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	4453	5024	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	3879	4452	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	intron	2217	3878	.	-	.
gb AC011536.6 gi 21617636	GIGOgene	exon	2015	2216	.	-	.

GIGOgene-version 1.1
date: Wed Mar 3 11:38:00 2004

gb AC006950.1 gi 4321125	GIGOgene	exon	38679	38875	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	37087	38678	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	36515	37086	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	35943	36514	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	35369	35942	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	33707	35368	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	33507	33706	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	30412	33506	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	30028	30411	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	27184	30027	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	26912	27183	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	intron	26788	26911	.	-	.
gb AC006950.1 gi 4321125	GIGOgene	exon	26622	26787	.	-	.

GIG0gene-version 1.1
 ## date: Wed Mar 3 11:38:33 2004

gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	121141	121203	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	121204	127460	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	127461	128708	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	128709	131034	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	131035	131384	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	131385	137124	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	137125	137695	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	137696	139977	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	139978	140582	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	140583	141509	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	141510	142077	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	142078	149443	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	149444	150036	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	150037	152825	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	152826	153420	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	153421	153538	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	153539	153738	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	153739	155613	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	155614	155755	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	155756	159202	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	159203	159540	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	159541	160850	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	160851	161424	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	161425	161818	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	161819	162435	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	162436	163163	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	163164	163707	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	163708	165277	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	165278	165870	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	165871	168763	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	168764	169358	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	169359	169476	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	169477	169676	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	169677	171903	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	171904	172045	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	172046	175457	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	175458	175795	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	175796	177105	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	177106	177679	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	177680	178073	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	178074	178690	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	178691	179414	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	179415	179958	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	179959	181529	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	181530	182122	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	182123	185015	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	185016	185610	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	185611	185728	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	185729	185928	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	185929	188154	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	188155	188296	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	188297	191708	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	191709	192046	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	192047	193364	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	193365	193938	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	193939	194326	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	194327	194943	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	194944	195669	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	195670	196213	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	196214	197805	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	197806	198377	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	198378	198949	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	198950	199523	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	199524	201185	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	201186	201385	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	201386	204480	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	204481	204864	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	204865	207768	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	207769	207980	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	intron	207981	208104	.	+	.
gb AC007842.1 gi 5080755+gb AC011536.6 gi 21617636+gb AC006950.1 gi 4321125	GIG0gene	exon	208105	208270	.	+	.

NM_003890.1	Exons	AC007842.1	AC011536.6	AC006950.1
1-63	1	52868-52806		
64-1311	2	46548-45301		
1312-1661	3	42974-42625		
1662-2232	4	36884-36314		
2233-2837	5	34031-33427		
2838-3405	6	32499-31932		
3406-3998	7	24565-23973		
3999-4593	8	21183-20589		
4594-4793	9	20470-20271		
4794-4935	10	18395-18254		
4936-5273	11	14806-14469		
5274-5847	12	13158-12585		
5848-6464	13	12190-11574		
6465-7008	14	10845-10302		

7009-7601	15	8731-8139	38124-37532	
7602-8196	16	5245-4651	34638-34044	
8197-8396	17	4532-4333	33925-33726	
8397-8538	18	2128-1987	31498-31357	
8539-8876	19	2128-1987	27944-27607	
8877-9450	20		26296-25723	
9451-10067	21		25328-24712	
10068-10611	22		23987-23444	
10612-11204	23		21872-21280	
11205-11799	24		18386-17792	
11800-11999	25		17673-17474	
12000-12141	26		15247-15106	
12142-12479	27		11693-11356	
12480-13053	28		10037-9464	
13054-13670	29		9075-8459	
13671-14214	30		7732-7189	38875-38679
14215-14786	31		5596-5025	37086-36515
14787-15360	32		4452-3879	35942-35369
15361-15560	33		2216-2015	33706-33507
15561-15944	34		2216-2015	30411-30028
15945-16216	35			27183-26912
16217-16383	36			26787-26622

We use General Feature Format (GFF) for sequences annotations, specified at http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml.

The column definition for the above four GFF annotation outputs is as follows:

<seqname> <source> <feature> <start> <end> <score> <strand> <frame>

<**seqname**> The name of the sequence. The seqname is the identifier of the sequence in an accompanying FASTA format file,

<**source**> The source of this feature, in our case the GIGOgene program,

<**feature**> The feature type name, either exon or intron,

<**start**> Start of the feature within the sequence,

<**end**> End of the feature within the sequence,

<**score**> Currently there is no score in our program, '.' goes instead,

<**strand**> One of '+', '-' or '.' indicating sense or antisense strand for DNA.

<**frame**> Currently there is no frame information in our program, '.' goes instead.

For test case we join few local annotations to make a global gene structure prediction, as indicated by structural layout at the end of the annotation.

In the layout, the first column is exon location in transcript, the second column is exon index and the following columns are exonic locations within genomic clones. Note, that once a certain layout location is on antisense strand, the start is bigger than the end.

2 Advanced use of the program

2.1 Brief program description

The program automatically filters out High Scoring Segment Pairs (HSPs) resulting from polyA tail, unless instructed not to.

Most of the parameters mentioned here can be adjusted in parameter file, described in subsection 2.2.

The program parses BLASTN output for HSPs, parameters of the query and target sequences. The `queryResult` object is a structure representing the result of the query, containing `BlastSequence` objects sorted by score. Each `BlastSequence` object contains set of `HighScoringPair` objects. Every `queryResult` and `BlastSequence` object contains unique sequence *name*, but no real sequences. There are three ways to obtain complete sequences of query (mRNA) and targets (DNA) by their corresponding names:

- From Entrez on the web by GI or Accession number only,
- From FASTA formatted file (every record has to have a unique FASTA name!!!),
- From database (we use JDBC connection to PostgreSQL database with sequences).

After we put the sequences in the `queryResult` object and to every valid `BlastSequence` (we require expectation to be $1e1$ with bit score higher 200.0), we form `IntervalNode` objects out of the `BlastSequence` objects. We split each `BlastSequence` object into two `IntervalNode` objects based on a hit strand of HSPs (Plus/Minus).

We convert all the DNA sequences into Positive strand, and we use java String indexing system for all the internal calculations. In order to extract subsequence from `IntervalNode`, we add one to the end internal coordinate of a subsequence, since all internal coordinates go as $[0, \dots, length - 1]$ for original $[1, \dots, length]$ range (see `java.lang.String` object documentation). DNA coordinates get converted back only at the very end of the calculations for display and storage purposes. All the debug information is presented in natural java `String` coordinates (positive strands only).

2.2 Parameters settings

There are numerous parameters in the file

`DIRECTORY/BlastObject/Parameters/clusterResultsParse.xml`. Here we briefly explain the parameters meaning. A more comprehensive coverage of parameter meaning could be found in [1].

Parameters are the following:

thresholde Cutoff value for expectation in BLASTN hit,

thresholdbitscore Cutoff value for bitscore in BLASTN hit,

biggest_graph_size Limitation for the interval graph size - it can not be bigger than this size,

debug_mode Whether we would like to see the debugging information on the screen,

draft_token Draft HTGs should be marked with this token in the FASTA description to be eliminated from further consideration,

version Version of the program,

hit_get_type The way we retrieve hit sequences, it could be FASTA, DATABASE, URL or NONE if we don't need to,

query_get_type The way we retrieve query sequences, it could be FASTA, DATABASE, URL or NONE if we don't need to,

blast_type This option is nucleic/protein,

query_fasta_file FASTA file with the query sequences to read from,

hit_fasta_file FASTA file with the target sequences to read from,

append Overwrite or append sequences to the output file,

db_username User name for JDBC connection,

db_password Password for JDBC connection,

db URL and Database name to connect to using JDBC connection,

allowed_gap Sometimes HSPs do not touch on transcript side. We need to allow certain unaligned nucleotides to fill this gap,

anchor_size Size of an Exon and Intron part to construct an anchor of,

fasta_width Width of the FASTA format to write in,

k Exponential weight function parameter,

exponent Exponential weight function parameter,

cut_off Cutoff average percentage for the exponential weight function,

d_unamb D parameter/disambiguation. There must be a reason for a break. Do not penalize it too much,

e_unamb E parameter/disambiguation. We do not penalize big gaps in DNA HSPs sequence. Some genes have a gap of size 35 or more confirmed by evidence,

url_text The way we start Entrez query to retrieve sequences with certain gi number from Entrez database. Needs to be complemented by `uid=GI#1&uid=GI#2&...&uid=GI#N`, where N is less than 20. Please refer to <http://www.ncbi.nlm.nih.gov/Entrez/tutor.html>,

url_search The way we start Entrez query to retrieve sequences by accession number from Entrez. Needs to be completed with `&term=accession#`. Please refer to <http://www.ncbi.nlm.nih.gov/Entrez/tutor.html>,

fasta_header All the FASTA names will start with this prefix in the output file,

d_alignment D -parameter for the spliced alignment,

e_alignment E -parameter for the spliced alignment,

punish_alignment Punish multiplier for the non-intron related break in DNA,

max_number_unamb_seq_per_bs Some hit results need help in stopping making unambiguous HSP sequences,

failure_counter The number of tries we do to get sequence from the web,

max_rna_gap The gap for small exons. Assume we have burst of small exons of size 10. The BLASTN word is of size 11, we add one and multiply by 10, which gives gap maximum gap size 120,

shortest_intron The shortest intron possible in this organism,

filter_poly_a Filter the PolyA tails and all HSPs resulting from of them,

global_write Write down the global gene structure prediction or the local predictions with no global result.

2.3 Elements in the output file

The GIGGene program writes output in FASTA format. The description of the sequences in output file contain xml-like *elements*, with text describing different parameters of the sequence.

The elements are the following:

origin The DNA clone from where the feature sequence originates,

feature_type Exon or Intron,

location_origin The segment location within the clone, containing the feature,

location_evidence The segment location within the transcript, aligning to feature in the genomic clone annotated,

strand Plus or Minus for the clone. Transcript is always on positive strand, while clone could be sequenced on sense or anti-sense strands,

evidence The transcript, alignment to which was the evidence for the feature,

feature_number Auxiliary information for parsers.

References

- [1] Alexandre Tchourbanov, Daniel Quest, Hesham Ali, Mark Pauley, and Robert Norgren, *A new approach for gene annotation using unambiguous sequence joining*, Proceedings of IEEE Computer Society Bioinformatics Conference, CSB03, IEEE Computer Society Press, Aug. 2003.