

NCBI Mini-Course

Unmasking Genes in Human DNA

Dr. Medha Bhagwat, NCBI
(bhagwat@ncbi.nlm.nih.gov)

Dr. David Wheeler, NCBI
(wheeler@ncbi.nlm.nih.gov)

NCBI Mini-Course

Unmasking Genes in Human DNA

(<http://www.ncbi.nlm.nih.gov/Class/wheeler/Javagene/gg.html>)

This course is an introduction to mining the human genome.

First, we will predict the exons in the protein-coding genes using two gene prediction tools:

1. GenScan (<http://genes.mit.edu/GENSCAN.html>)
2. GeneMark (<http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi>)

in conjunction with

1. BLASTX (<http://www.ncbi.nlm.nih.gov/BLAST/>)
2. BLASTN against EST database. (<http://www.ncbi.nlm.nih.gov/BLAST/>)
3. Comparison with the mouse and rat genomic sequences (<http://www.ncbi.nlm.nih.gov/BLAST/>)

In addition, we will also predict the presence of

1. Promoters using PROSCAN (<http://bimas.dcrf.nih.gov/molbio/proscan/>)
2. Repeat elements using RepeatMasker (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>)

We will then assemble the amino acid sequence of the gene product, on the basis of the exons chosen, using the translation tool provided on the class web page.

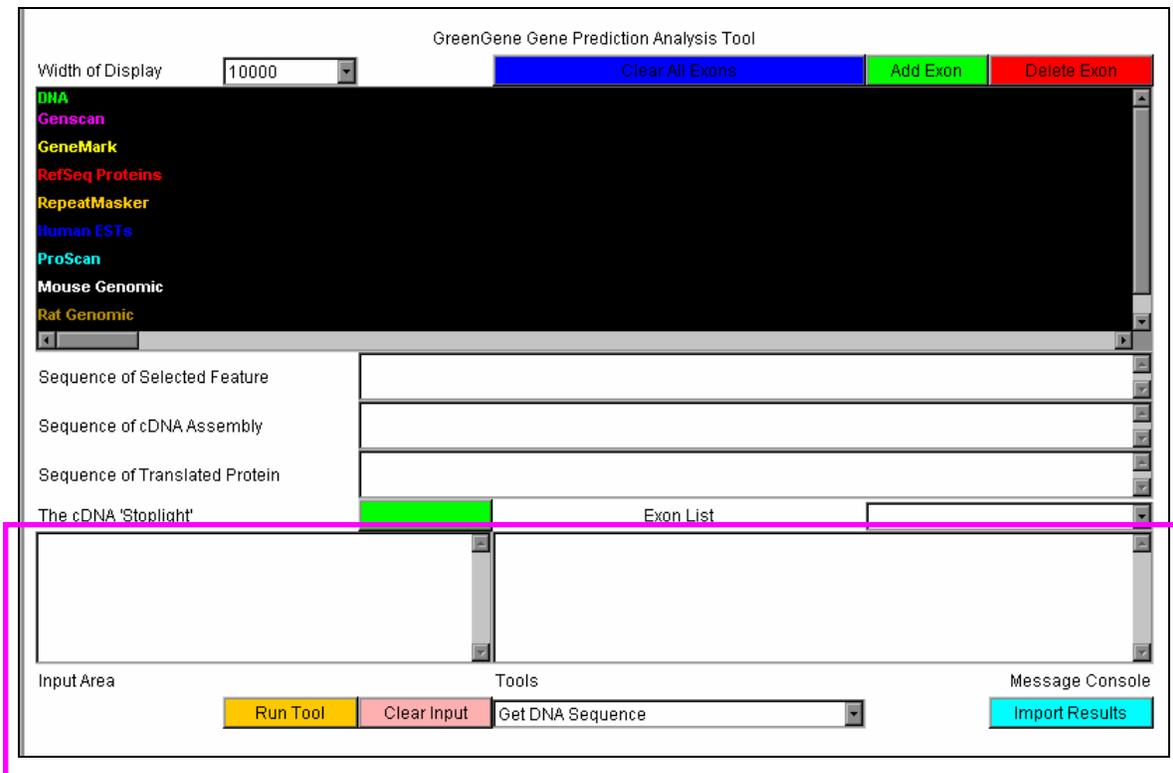
The web page contains links to all the necessary tools for the analysis. It also includes an ability to translate the DNA sequence into amino acid sequence by selecting the appropriate exon sequences. During the first hour, an instructor will walk you through an analysis of some genomic sequences. During the second hour of the class, you will perform the same analysis using different genomic sequences that will be provided to you.

Greengene is a Java application which accesses several web-based sequence-analysis tools relevant to the identification of eukaryotic genes, then captures and integrates their output in a single view for ease of comprehension. Using **Greengene**, exons can be picked interactively and assembled into a coding sequence, then translated into a protein product. The exon choices made by the user reflect the information provided by several of the tools used rather than that of a single tool and are, therefore, more reliable. The tools accessed are GenScan and Genemark (exon prediction), Repeatmasker (repeat identification), Proscan (promoter prediction), and blastx and blastn (to support exon prediction).

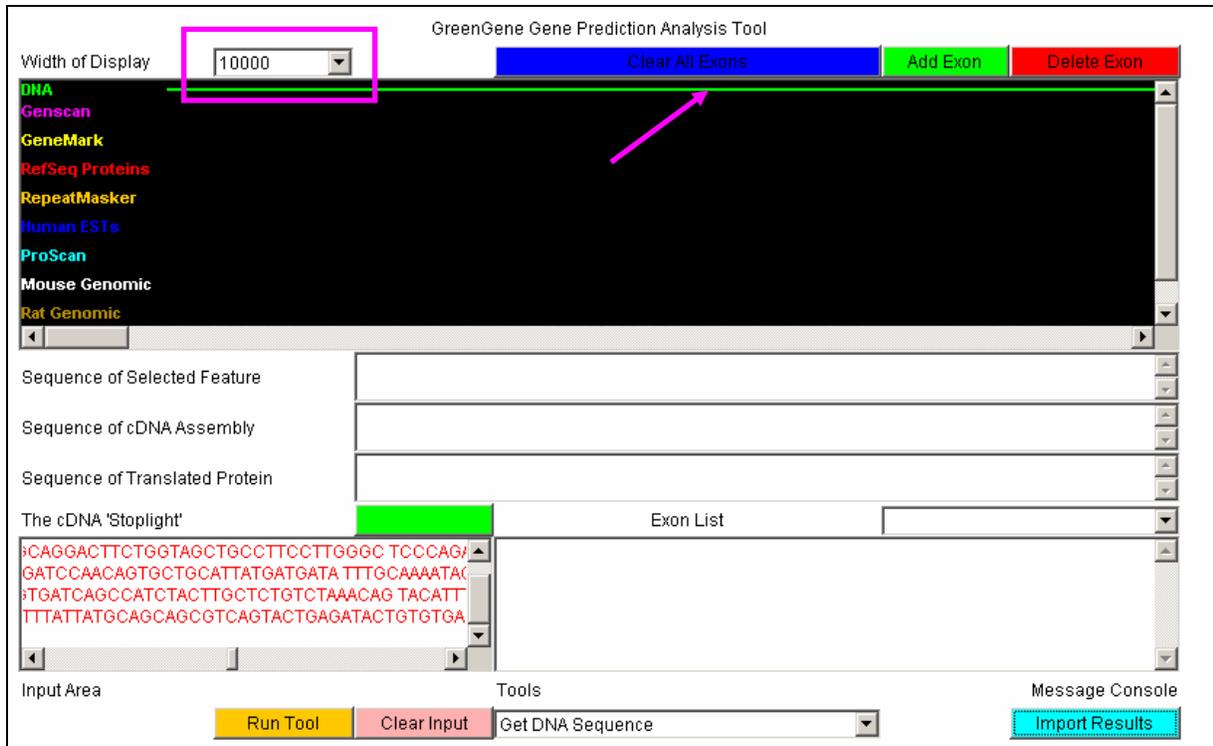
At least 50% of the human genome sequence is made up of repeat elements. Most of the eukaryotic protein coding genes contain exons and non-coding introns. The exons make up only about 1% of the human genome.

To correctly identify the exons in a eukaryotic gene you will need to compare the output of at least two different gene-prediction programs and couple this with the outputs given by a blastx search against protein sequences and a blastn search against ESTs. Greengene performs the data integration for you. The most reliable exons should be predicted by both programs and should align with blastx and/or blastn hits. In some cases, an exon prediction program may generate a potentially spurious exon without blastx or blastn support. In this case, you may want to exclude this exon in your gene model assembly. **Greengene** will allow you to do this easily so that you can create a custom coding sequence and the corresponding protein sequence without having to accept the automatically generated output of any single program.

Sequence-analysis tools are selected from a list-box labeled “Tools”. Once selected, a tool can either be “Run” or its output can be “Imported” into **Greengene**. When a tool is run, a second browser window is opened to the input page of the tool. The sequence to be analyzed must then be pasted into the tool’s input box and the tool must be run. When the tool has returned its output, the entire output should be selected (Select All), and pasted in the “Input Area”. To import the data, the “Import Results” button is then pressed.

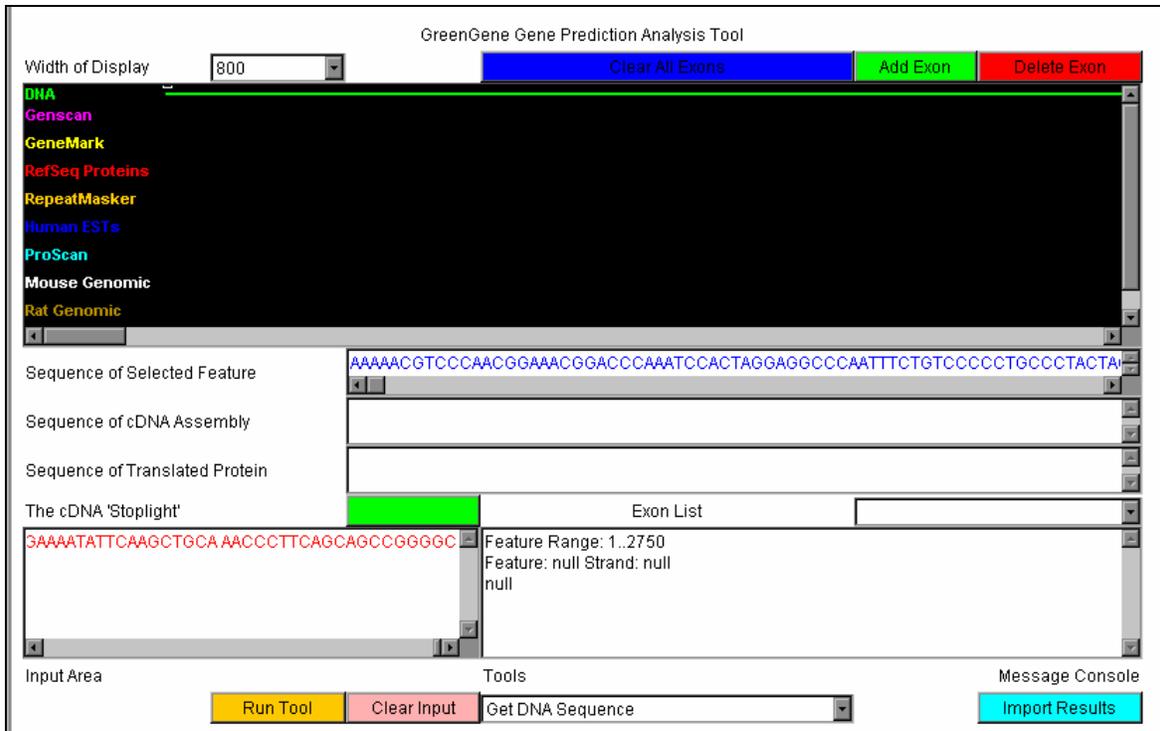


To import one of the example sequences, select the “Sequence” tool, click on “Run”, highlight and copy the sequence desired, paste it in the “Input Area” and click on “Import Results.” You may paste DNA sequence of your choice not provided in the Sequence file.



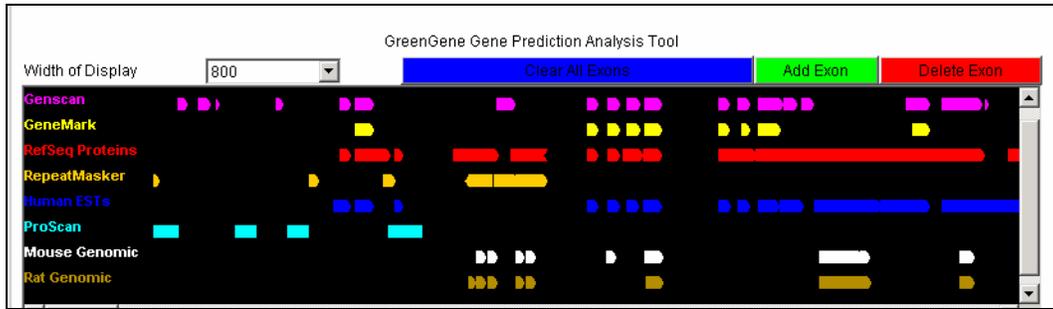
Once the import operation is complete, colored blocks will appear on the line corresponding to the proper tool in the display pane under the green line representing the DNA sequence. These blocks give the locations of features returned by the analysis tool. The width of the Display panel can be adjusted using the list box. Select 800 to zoom out to the full length of the DNA sequence.

To get information about a feature, click on the colored block. A click prints information in the “Message Console” area and puts the sequence of the feature in question into the “Sequence of Selected Feature” area. To put the entire DNA sequence into the “Sequence of Selected Feature” area, as required for pasting into an analysis tool, just click the green sequence line.

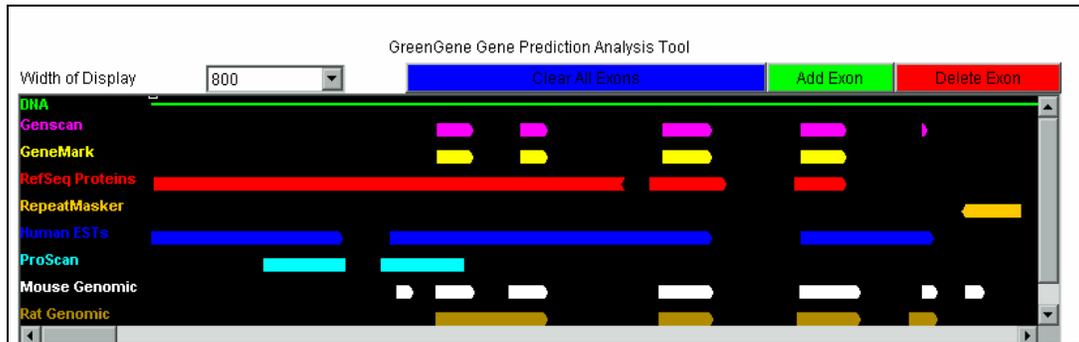


To assemble a coding sequence, click on the first exon you want to include from the GenScan or the GeneMark output and press the “Add Exon” button. Continue doing this until you have added the sequences of all the exons you wish to include in your gene model. The cDNA sequence of the assembled gene model is shown in the “Sequence of cDNA Assembly” box and its amino acid translation in the first reading frame in the “Sequence of Translated Protein”. Look out for asterisks as they indicate stops and should be present only at the end of your protein translation. Internal stops indicate a problem in your exon assembly and are indicated by red color in “The cDNA stoplight” whereas a terminal stop codon is indicated by grey color. You can get to the list of all exons used in the assembly from the “Exon List” pull down menu. You may delete an exon by selecting it from the list and clicking on “Delete Exon. There is also a button to “Clear All Exons” and start over the assembly.

Below are **Greengene** outputs for 2 example DNA sequences that will be used in the class. To analyze one of these yourself, select “Initial Sequence” and hit ‘Run’. This will open a browser window containing the initial sequences to use as well as complete outputs from the various tools to use in case of web problems. We will use DNA2 during the demonstration; run the entire analysis on DNA1 yourself.



DNA1: Note the 7th feature in the Genscan track; is this likely to be a real exon?



DNA2: All exon predictions supported by blastx and blastn hits.


```

GENSCAN 1.0      Date run: 22-Sep-104      Time: 13:26:04

Sequence 13:26:04 : 2750 bp : 57.42% C+G : Isochore 4 (57 - 100 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
-----
1.01 Intr +   862   960   99  0  0   84   51  168 0.956 14.17
1.02 Intr +  1113  1184   72  2  0   91   76  108 0.978 10.65
1.03 Intr +  1542  1679  138  2  0  107   61  184 0.998 19.44
1.04 Term +  1958  2086  129  1  0   78   48   92 0.906  3.32
1.05 PlyA +  2323  2328    6                1.05

```

GreenGene Gene Prediction Analysis Tool

Width of Display: 10000 Clear All Exons Add Exon Delete Exon

cDNA Model
Genscan
GeneMark
Blastx
RepeateMasker
Blastn
ProScan
Mouse
Rat

Sequence of Selected Feature: ACCATCTTCCCAATTCGGTTCAGAAAATATTCAGCTGCAAAACCCTTCAGCAGCCGGGGC

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight' Exon List

exons. It has been shown that predicted exons with higher probabilities are more likely to be correct than those with lower probabilities.

Feature Range: 1..2750
 Feature: null Strand: null
 hi there
 <?xml version="1.0"?><!DOCTYPE eSummaryResult PUBLIC "-//NLM/DTD

Input Area Tools: Genscan Message Console

Run Tool Import Results

GeneMark.hmm (Version 2.2a)
 Sequence name: Fri Feb 9 11:30:34 EST 2007
 Sequence length: 2750 bp
 G+C content: 57.42%
 Matrix: Homo sapiens
 Fri Feb 9 11:30:34 2007

Predicted genes/exons

Gene #	Exon #	Strand	Exon Type	Exon Range	Exon Length	Start/End Frame
1	1	+	Internal	862 960	99	1 3
1	2	+	Internal	1113 1184	72	1 3
1	3	+	Internal	1542 1679	138	1 3
1	4	+	Terminal	1958 2086	129	1 3

GreenGene Gene Prediction Analysis Tool

Width of Display:

DNA

GeneScan █ █ █ █ █

GeneMark █ █ █ █

RefSeq Proteins

RepeatMasker

Human ESTs

ProScan

Mouse Genomic

Rat Genomic

Sequence of Selected Feature:

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight' █ Exon List:

Feature Range: 1..2750
 Feature: null Strand: null
 null

Improvements or problems to the web page maintainer.

Input Area: Tools:

Message Console

RepeatMasker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.

Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.1.6

[Check Current Queue Status](#)

Basic Options

or

Sequence:

Search Engine: cross_match wublast

Speed/Sensitivity: rush quick default slow

DNA source:

Return Format: html tar file

Return Method: html email

Select a sequence file to process or paste the sequence(s) in FASTA format. Large sequences will be queued, and may take a while to process.

Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than WUBlast.

Select the sensitivity of your search. The more sensitive the longer the processing time.

Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.

Select the format for the results of your search. The "tar" option will return the results as a compressed archive file, and "html" will present the results as a summary web page with links to the individual data files.

The "HTML" return method will run RepeatMasker on your sequence and return the results immediately to your web browser, provided your sequences are short. The "email" return method will email you when your results are ready.

Return Method: html email

RepeatMasker on your sequence and return the results immediately to your web browser, provided your sequences are short. The "email" return method will email you when your results are ready.

RepeatMasker started 08-Feb-2007 07:55:21 PST

RepeatMasker version open-3.1.6
Search engine: WUBlast
analyzing file /usr/local/rmsrver/tmp/RM2sequpload_1170950111

Results

Right-click and select "Save As" to save results to your computer or click on the link to view the file in the browser.
Annotation File: [RM2sequpload_1170950111.out](#)
Masked File: [RM2sequpload_1170950111.masked](#)

SW score	perc div.	perc del.	perc ins.	query sequence	position begin	position end	in query (left)	matching repeat	position in repeat				
					begin	end	(left)	repeat	class/family	begin	end	(left)	ID
607	17.7	11.2	0.5	UnnamedSequence	2453	2622	(128)	C	MER20	DNA/MER1_type	(2)	217	30 1

GreenGene Gene Prediction Analysis Tool

Width of Display:

DNA

GeneScan ▶

GeneMark ▶

RefSeq Proteins ▶

RepeatMasker ▶

Human ESTs

ProScan

Mouse Genomic

Rat Genomic

Sequence of Selected Feature:

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight':

Exon List

Feature Range: 1..2750
Feature: null Strand: null
null

Web Promoter Scan Service - Mozilla Firefox

http://bimas.dort.nih.gov/molbio/proscan/


BioInformatics and Molecular Analysis Section
 Computational Bioscience and Engineering Lab, Division of Computational Bioscience
 Center for Information Technology, National Institutes of Health

WWW Promoter Scan

Function: Predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences.
 The **analysis** is done using the PROSCAN Version 1.7 suite of programs developed by [Dr. Dan Prestidge](#). Information on PROSCAN, including details on obtaining a copy, is maintained at the [Advanced Biosciences Computing Center](#), University of Minnesota.

A DNA sequence is all that needs to be supplied. There are no optional parameters for PROSCAN.

Please enter or paste a Nucleic Acid sequence to analyze (most [formats](#) accepted):

Echo input sequence (generally [recommended](#))

Be Forewarned!

Patience is a virtue: Analysis for a 10Kbp sequence may take as long as 5 minutes (or more)!

Credits: WWW implementation by [BIMAS staff](#)

Significant Signals:

Name	TFD #	Strand	Location	Weight
HSV_IE_repeat	S01565	-	351	1.363000
Sp1	S01542	-	354	3.608000
JCV_repeated_sequenc	S01193	-	364	1.658000
T-Åg	S00974	+	398	1.086000
NF-kB	S01498	-	432	1.008000
EARLY-SEQ1	S01081	+	472	6.322000
Sp1	S00801	+	473	2.755000
Sp1	S00802	+	474	3.292000
Sp1	S00781	-	478	2.772000
Sp1	S00978	-	479	3.361000
JCV_repeated_sequenc	S01193	-	479	1.658000
Sp1	S00952	+	490	50.000000
Sp1	S01542	-	499	3.608000
beta-pol_CS	S00559	+	510	8.603000
ATF	S01059	+	511	1.157000
CREB	S00969	+	511	3.442000
CREB	S00072	+	511	8.603000

GreenGene Gene Prediction Analysis Tool

Width of Display: Clear All Exons Add Exon Delete Exon

Sequence of Selected Feature: `GGTGCAGGACCATCTTCCCAATTCGGTTCAGAAAAATTTCAAGCTGCAAACCCTTCAGCAGCCGGGGC`

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight': Exon List:

Sp1	+	694	5.934000	Feature Range: 1..2750 Feature: null Strand: null null
Sp1	+	694	17.211000	

GreenGene Gene Prediction Analysis Tool

Width of Display: Clear All Exons Add Exon Delete Exon

DNA

Genscan █ █ █ █

GeneMark █ █ █ █

RefSeq Proteins █ █ █

RepeatMasker █

Human ESTs

ProScan █ █

Mouse Genomic

Rat Genomic

Sequence of Selected Feature: `GGTGCAGGACCATCTTCCAATTCCGTTTCAGAAAATATTC AAGCTGCAAACCTTCAGCAGCCGGGGC`

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight' █ Exon List

1_21789	gi 121537830 ref ZP_01669606.1 27.1	Feature Range: 1..2750
1_21789	gi 121537830 ref ZP_01669606.1 25.6	Feature: null Strand: null
1_21789	gi 39940408 ref XP_359741.1 29.6	null

NCBI **megablast BLAST**

Nucleotide Protein Translations Retrieve results for an RID

[What is Mega BLAST?](#)

[Search](#)

Load query file from disk Browse...

[Set subsequence](#) From: To:

[Choose database](#)

Human genomic plus transcript
 Mouse genomic plus transcript
 Others (nr, etc.):

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

```

# BLASTN 2.2.15 [Oct-15-2006]
# Query:
# Database: est_human
# Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end
# 100 hits found
1_10471 gi|18808711|gb|BM562537.1| 95.54 920 15 24 747 1642 1 918
1_10471 gi|80793350|dbj|DA433015.1| 98.92 555 2 4 722 1274 1 553
1_10471 gi|19808382|gb|BQ049042.1| 98.57 558 0 8 2 559 561 12
1_10471 gi|19122787|gb|BM805964.1| 97.86 514 5 6 829 1338 9 520
1_10471 gi|83211093|dbj|DB052544.1| 98.21 503 0 9 2 503 495 1
1_10471 gi|45711089|emb|AL535211.3| 98.35 485 0 8 2 486 477 1
1_10471 gi|83212966|dbj|DB281481.1| 98.14 485 0 9 2 485 477 1
1_10471 gi|43442377|emb|BX956949.1| 97.95 488 1 9 2 488 481 2
1_10471 gi|90833951|dbj|DB462532.1| 97.92 481 2 8 2 482 473 1
1_10471 gi|83053300|dbj|DA952946.1| 97.52 483 2 10 2 482 475 1
1_10471 gi|81127696|dbj|DA532865.1| 98.08 469 1 8 2 470 461 1
1_10471 gi|47378328|dbj|CN300733.1| 98.25 461 0 8 2 462 458 6

```

GreenGene Gene Prediction Analysis Tool

Width of Display:

Sequence of Selected Feature:

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Spotlight' Exon List

1_10471	gi 18981231 gb BM671333.1	100	Feature Range: 1..2750
1_10471	gi 18981231 gb BM671333.1	100	Feature: null Strand: null
1_10471	gi 18969574 gb BM663873.1	100	null

[What is discontiguous Mega BLAST?](#)

[Search](#)

```
GGACCATCTTCCCAATTCCGTTTCAGAAAAATATTCAAGCTGCAAAACCTTCAGCAGCCGGGGC
```

Load query file from disk

[Set subsequence](#) From: To:

[Choose database](#) Human genomic plus transcript
 Mouse genomic plus transcript

Others (nr etc.):

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Return alignment endpoints only

Now: or

Options for advanced blasting

[Limit by entrez query](#) AND

GreenGene Gene Prediction Analysis Tool

Width of Display:

Sequence of Selected Feature: **GGTGCAGGACCATCTTCCCAATTCCGTTTCAGAAAATATTC AAGCTGCAAACCTTCAGCAGCCGGGGC**

Sequence of cDNA Assembly:

Sequence of Translated Protein:

The cDNA 'Stoplight':

Feature ID	Gene/Transcript	Exon List
1_25007	gi 83274085 ref AC_000033.1 AC_000	Feature Range: 1..2750
1_25007	gi 94471499 ref NC_000079.4 NC_000	Feature: null Strand: null
1_25007	gi 83274089 ref AC_000035.1 AC_000	null

Input Area: Tools: Message Console:

Choose database:

- Human genomic plus transcript
- Mouse genomic plus transcript
- Others (nr etc.):

Return alignment endpoints only:

Now: or

NEW. Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Options for advanced blasting

Limit by [entrez query](#) AND

GreenGene Gene Prediction Analysis Tool

Width of Display: 800

Buttons: Clear All Exons, Add Exon, Delete Exon

Genomic tracks: DNA, Genscan, GeneMark, RefSeq Proteins, RepeatMasker, Human ESTs, ProScan, Mouse Genomic, Rat Genomic

Sequence of Selected Feature: GGTGCAGGACCATCTTCCCAATTCCGTTTCAGAAAATATTC AAGCTGCAAACCTTCAGCAGCCGGGGC

Sequence of cDNA Assembly: [Empty]

Sequence of Translated Protein: [Empty]

The cDNA 'Stoplight': [Green]

Exon List: [Empty]

Feature Range: 1..2750
Feature: null Strand: null
null

Input Area: [Empty]

Tools: Run Tool, Clear Input, Rat Genomic (X-Species Megablast)

Message Console: Import Results

Create a custom coding sequence and the corresponding protein sequence:

GreenGene Gene Prediction Analysis Tool

Width of Display: 800

Buttons: Clear All Exons, Add Exon, Delete Exon

Genomic tracks: DNA, Genscan, GeneMark, RefSeq Proteins, RepeatMasker, Human ESTs, ProScan, Mouse Genomic, Rat Genomic

Sequence of Selected Feature: GAGTATAAGCTGATGTACGGGATGCTCTTCTCTATCCGCTCGTTTGT CAGCAAGATGTCCCCGCTAGAC

Sequence of cDNA Assembly: TATAAGCTGATGTACGGGATGCTCTTCTCTATCCGCTCGTTTGT CAGCAAGATGTCCCCGCTAGACATG

Sequence of Translated Protein: MTVHNLYLFDNRNGVCLHYSEWHRKKQAGIPKEEYKLMYGLFSIRSFVSKMSPLDM

The cDNA 'Stoplight': [Green]

Exon List: 1.02 Intr + 1113 2 3

Feature Range: 1113..1184
Feature: 1.02 Intr + 1113 1184 72 2 0 91 76 108 0.978 10.65
Strand: +
null

GreenGene Gene Prediction Analysis Tool

Width of Display: 800

Buttons: Clear All Exons (blue), Add Exon (green), Delete Exon (red)

Genomic tracks (from top to bottom):

- DNA: Green line with yellow boxes representing exons.
- GeneScan: Yellow boxes representing predicted exons.
- RefSeq Proteins: Red arrows representing protein models.
- RepeatMasker: Yellow bars representing masked repeats.
- Human ESTs: Blue bars representing ESTs.
- ProScan: Cyan bars representing predicted exons.
- Mouse Genomic: White arrows representing mouse gene models.
- Rat Genomic: Yellow arrows representing rat gene models.

Sequence of Selected Feature: AATAAA

Sequence of cDNA Assembly: TCGCTCCCGACTGGACTCCTATGTTTCGCTCTCTGCCCTTCTTCCGCCCGGGCTGGCTGAAATAAA

Sequence of Translated Protein: PTGIKVMNTDLGVGPIRDVLHHIYSALYVELWKNPLCPLGQTVQSELFRSRLDSYVRSLPFFSARAG*NK

The cDNA 'Stoplight'

Exon List	1.05 PlyA + 2323 2 9
1_3122_gi 62750810 ref NC_005109.2 NC_005109.79..	Feature Range: 2323..2326
1_3122_gi 109658149 ref AC_000089.1 AC_00008986..	Feature: 1.05 PlyA + 2323 2328 6 1.05
1_3122_gi 62750821 ref NC_005120.2 NC_005120.86..	Strand: +
	null