NCBI Mini-Course

Unmasking Genes in Human DNA

By

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.dcrt.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.

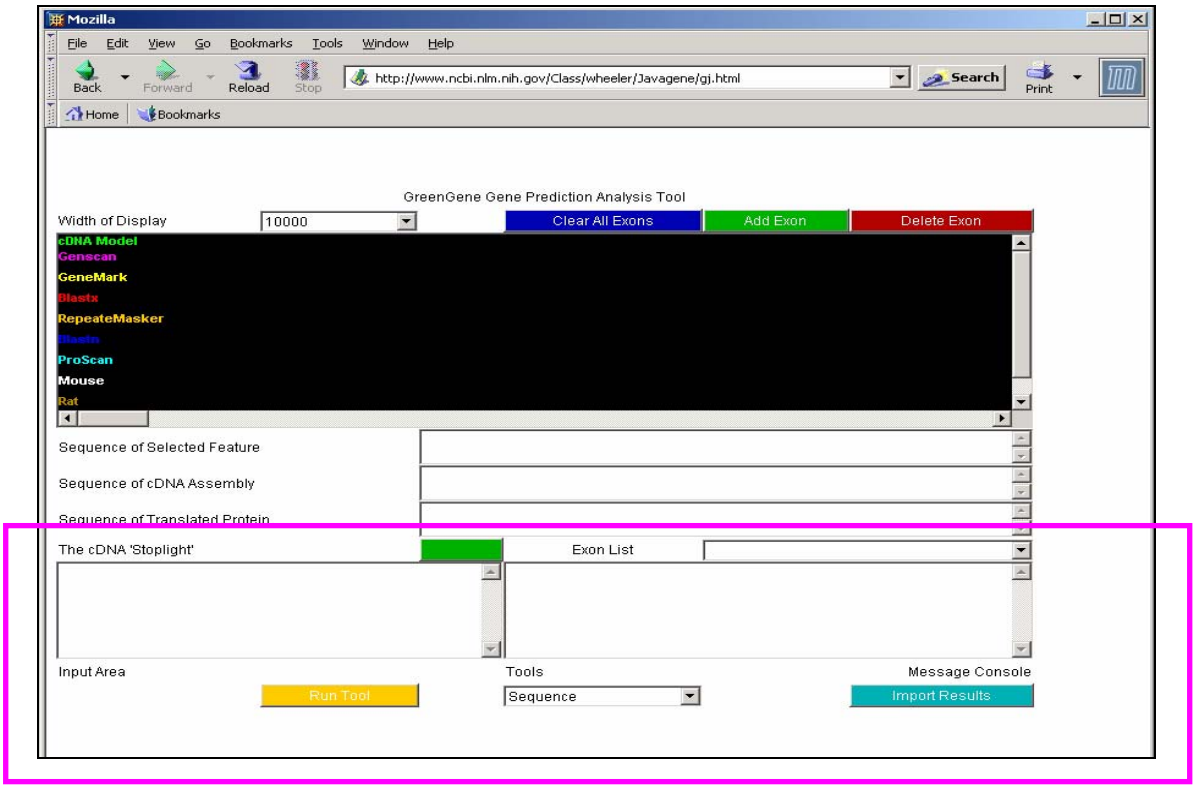
Greengene is a Java application which accesses several web-based sequence-analysis tools relevant to the identification of eukaryotic genes, then captures and integrates their output in a single view for ease of comprehension. Using **Greengene**, exons can be picked interactively and assembled into 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

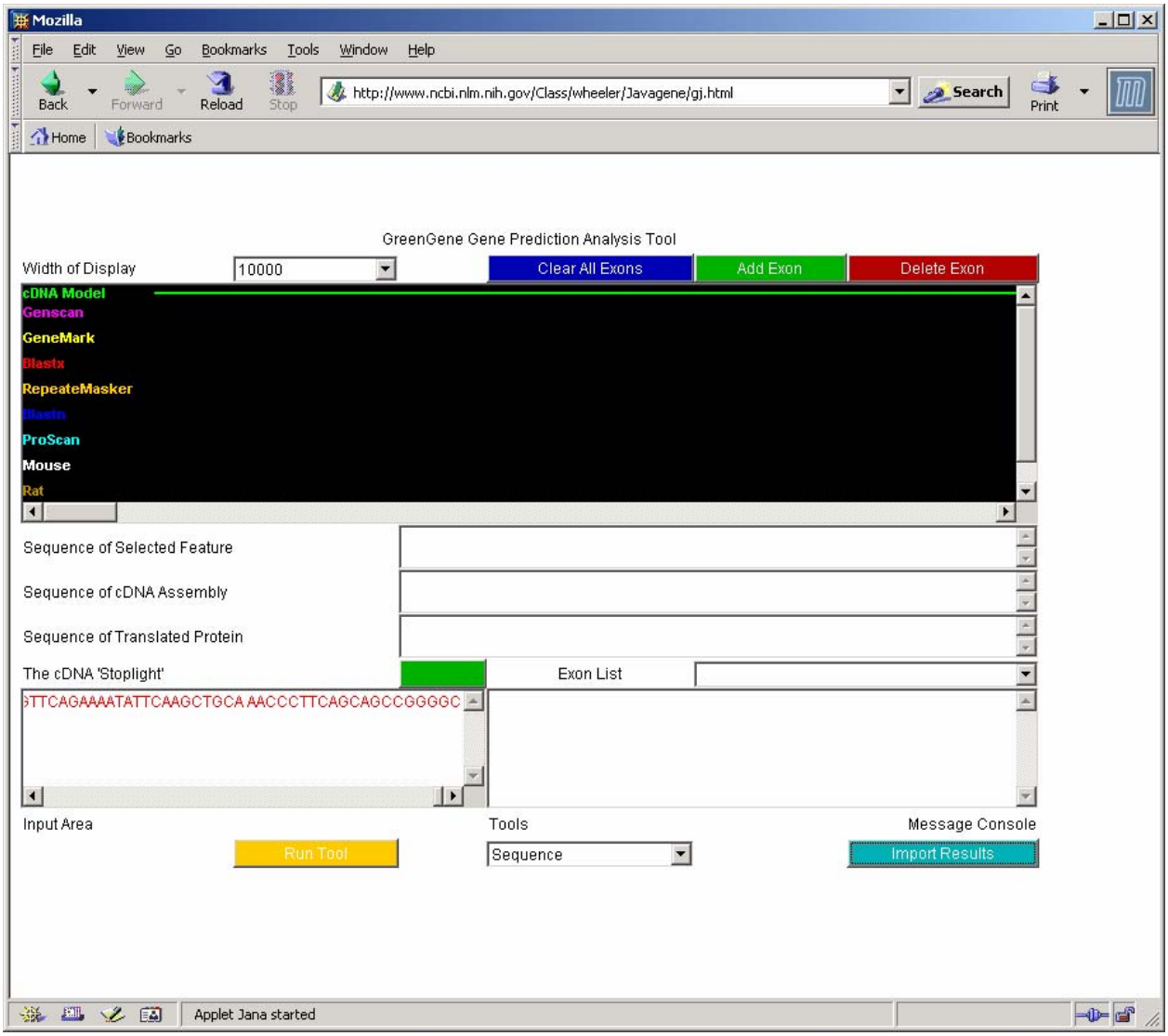
make up only about 1% of the human genome.

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.

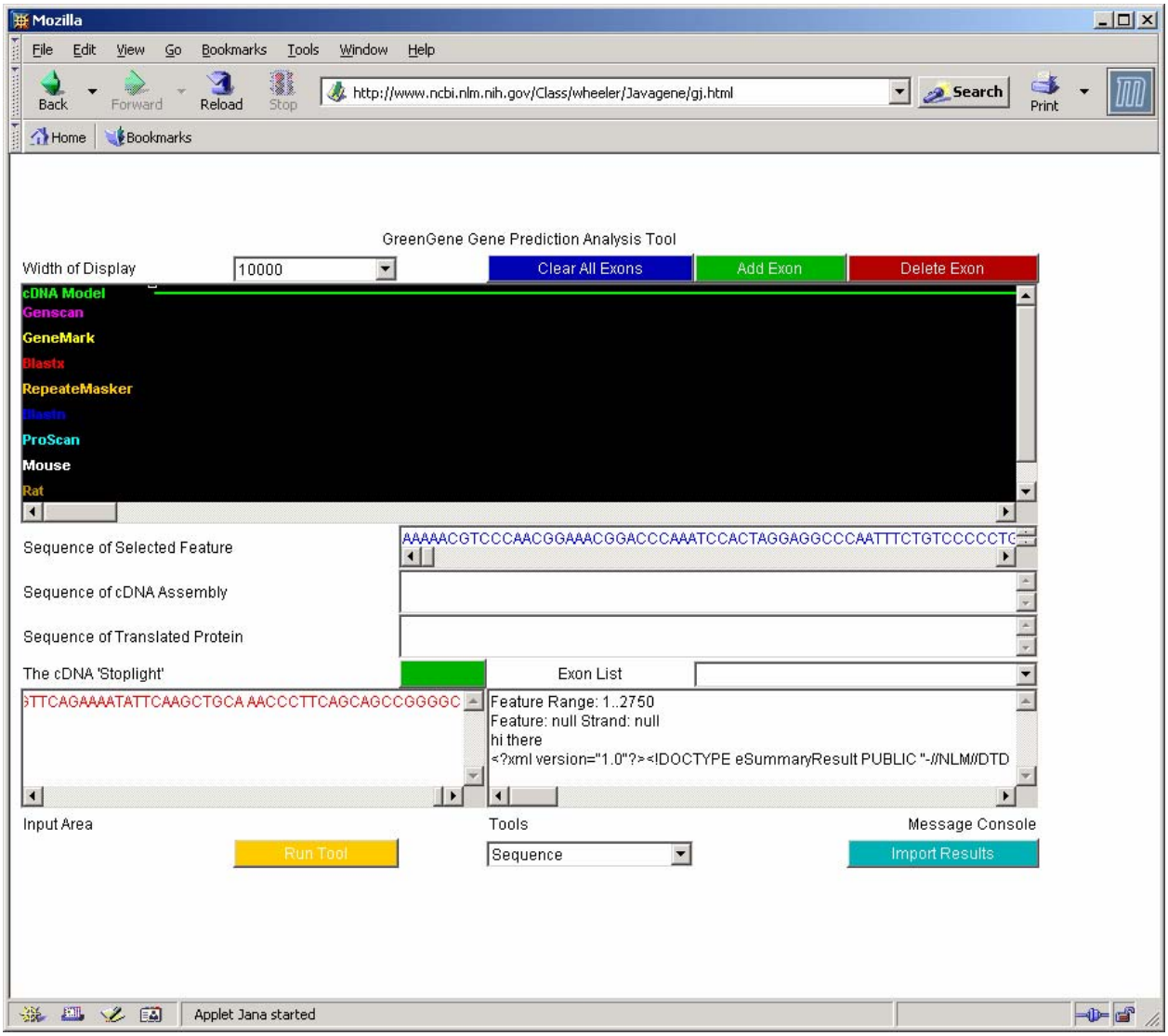


To import one of the example sequences, select the “Sequence” tool, click on “Run”, highlight and copy the sequence desired, paste it in the “Input Area” and click on “Import Results.” You may paste DNA sequence of your choice not provided in the Sequence file.



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.

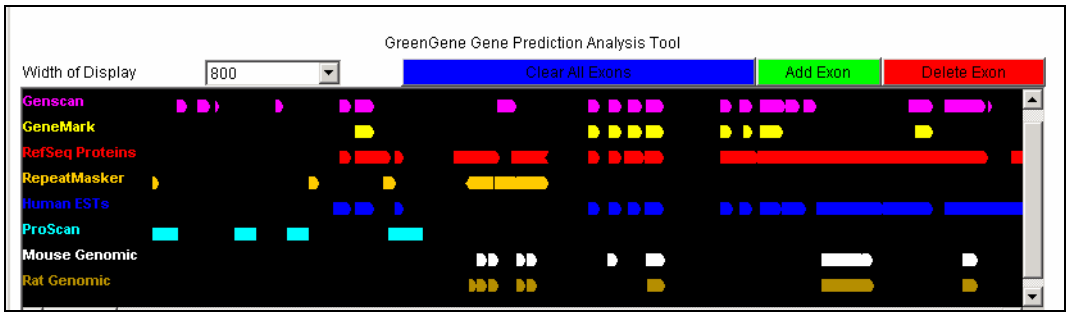
To get information about a feature, click on the colored block. A click prints information in the “Message Console” area and puts the sequence of the feature in question into the “Sequence of Selected Feature” area. To put the entire DNA sequence into the “Sequence of Selected Feature” area, as required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



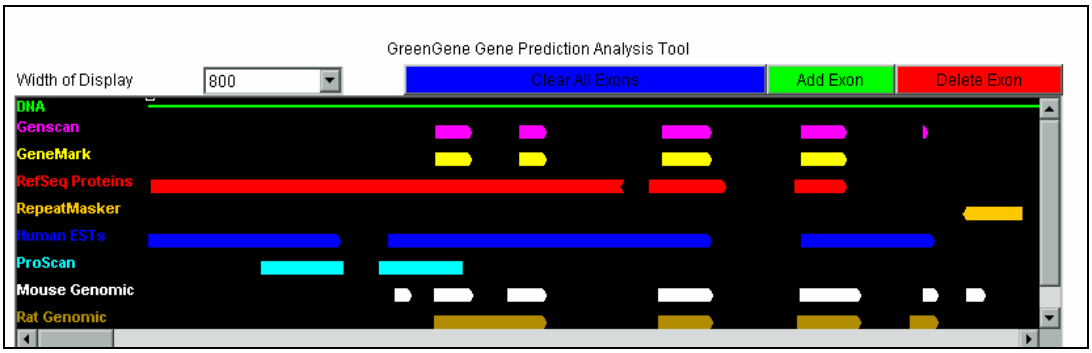
To assemble a coding sequence, click on the first exon you want to include from the GenScan or the GeneMark output and press the “Add Exon” button. Continue doing this until you have added the sequences of all the exons you wish to include in your gene model. The cDNA sequence of the assembled gene model 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. 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.

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

```
              \\|//
              (o o)
-. .-.   .-oOOo-(_)-oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.
||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\
|/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-'
```

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

News:

- **The server is back online. Apologies for the delay.**
  - The interface now supports file upload, allowing you to input long sequences without cutting and pasting. The old cut-and-paste method also works.
  - Also, the server now outputs PDF (portable document format) images of predicted gene locations, as well as Postscript.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored): Browse

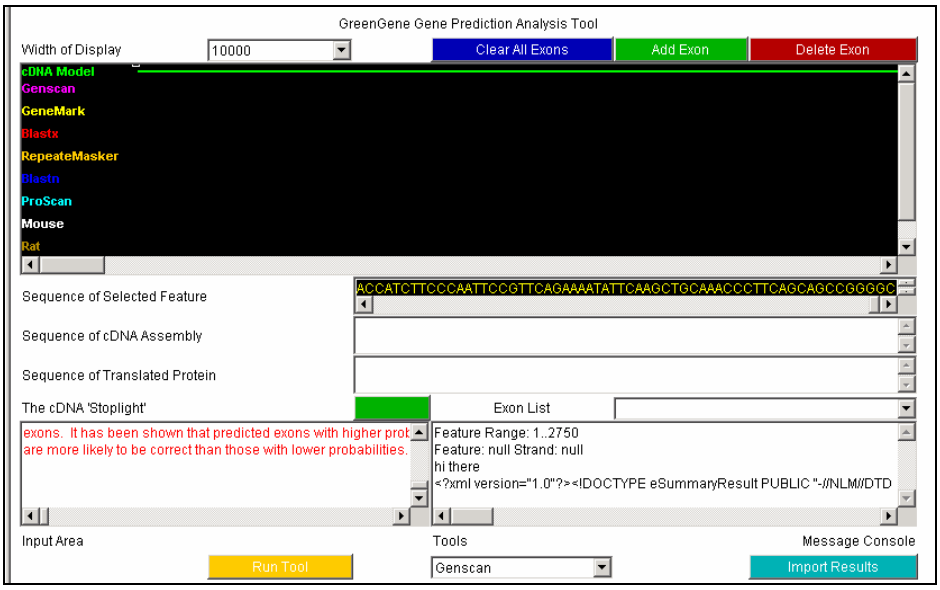
Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

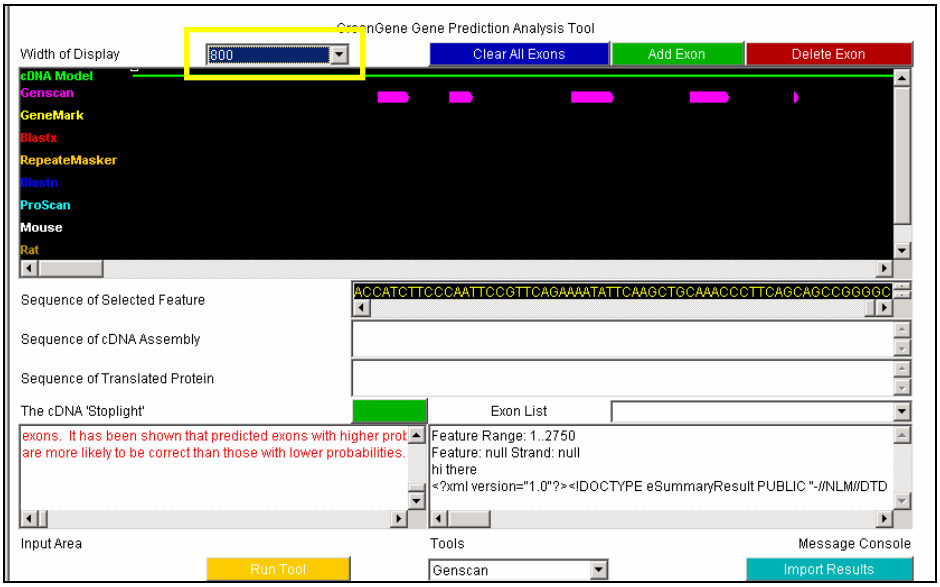
```
TCCCAAAGGTGCAGGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
```

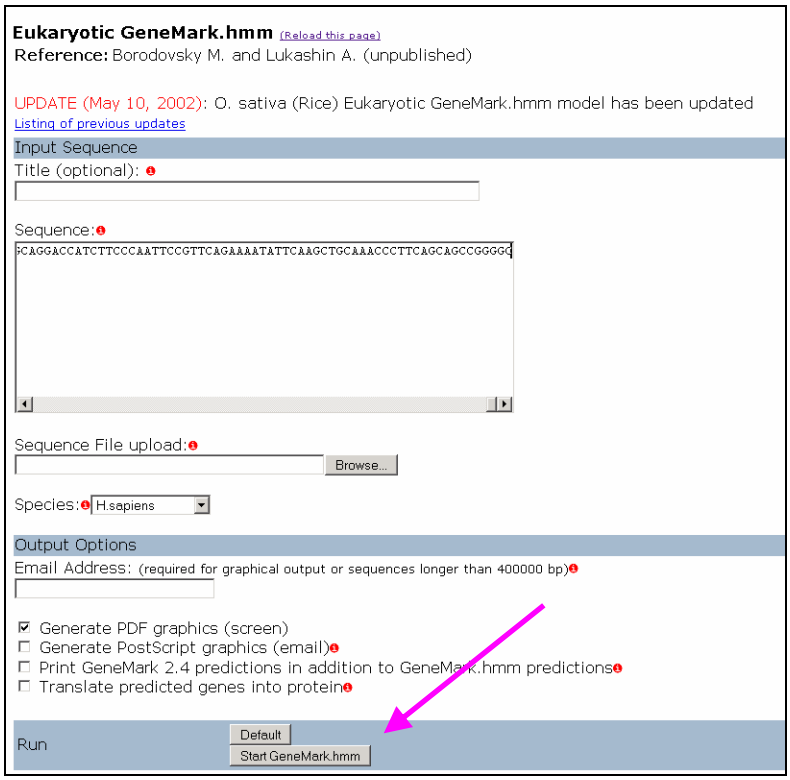
To have the results mailed to you, enter your email address here (optional):

Run GENSCAN Clear Input

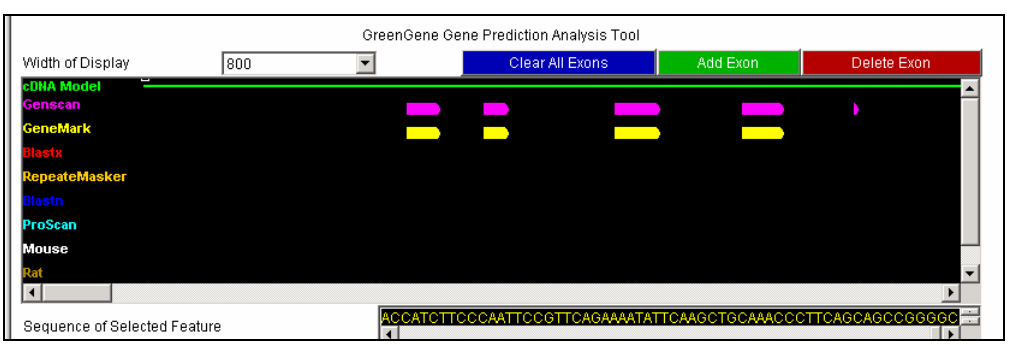
```
GENSCAN 1.0      Date run: 22-Sep-104     Time: 13:26:04

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
```








```
GeneMark.hmm (Version 2.2a)
Sequence name: Wed Sep 22 13:31:04 EDT 2004
Sequence length: 2750 bp
G+C content: 57.42%
Matrix: Homo sapiens
Wed Sep 22 13:31:05 2004

Predicted genes/exons

Gene Exon Strand Exon          Exon Range     Exon      Start/End
  #    #         Type                         Length       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
```



INSTITUTE FOR Systems Biology

RepeatMasker Web Server

RepeatMasker screens DNA sequences in fasta format against a library of repetitive elements and returns a mas
searches as well as a table annotating the masked regions.
Reference: A.F.A. Smit & P. Green, unpublished data. Current Version: 3.0.2

Check Current Queue Status

Basic Options

Large sequences will be queued, and may take a while to process.
Enter the file to process: [ ] Browse...

Or paste the sequence(s) in FASTA format:

```
GGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC
```

Select return format: (•) html ( ) tar file ( ) links
Select return method: (•) html ( ) email [Your email address]
[Reset] [Submit Sequence]

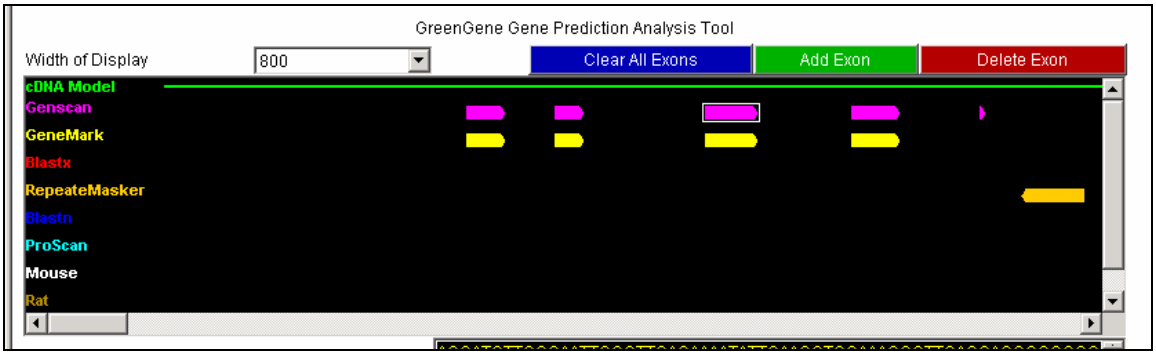
Advanced Options

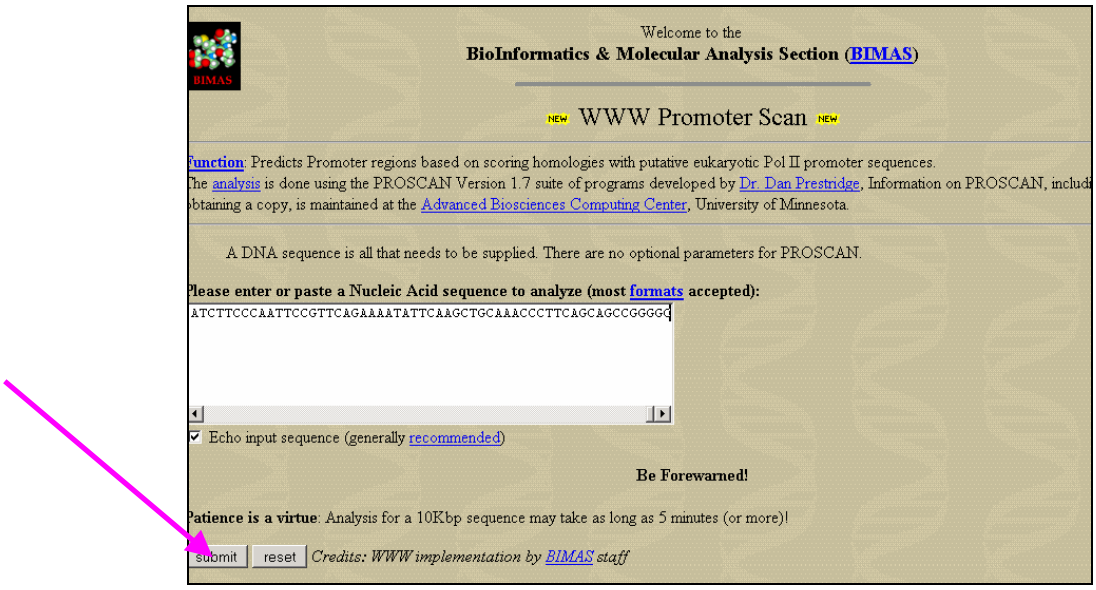
Speed/Sensitivity: (•) rush ( ) quick ( ) default ( ) slow

DNA source: [Human]

Repeat Annotations:

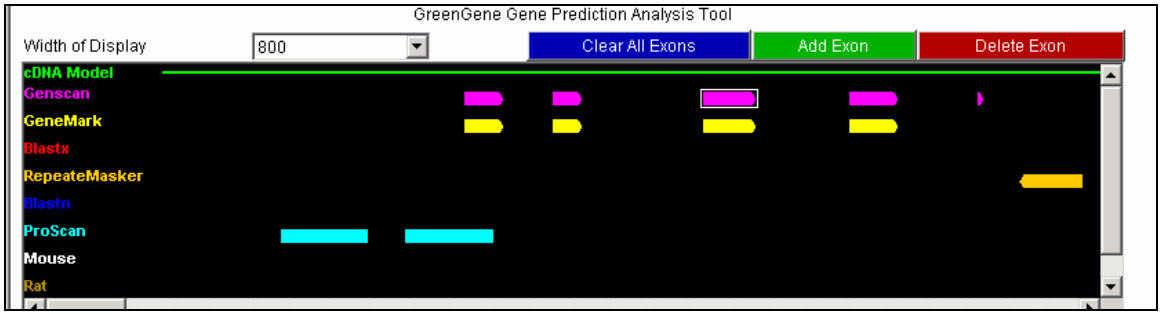
SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)		matching repeat repeat	class/family	position in repeat begin	end	(left)	ID
618	17.8	11.2	0.6	UnnamedSequence	2453	2622	(128)	C	MER20	DNA/MER1_type	(2)	217	30	1





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-Ag	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
CREB	S02107	+	511	3.886000
ATF	S01940	+	511	3.721000
CREB	S00144	+	512	1.912000
E4F1	S01252	+	513	3.824000
E4F1	S01249	+	513	3.764000
ATF/CREB	S00534	+	513	1.564000
CREB	S00489	-	515	1.147000
ATF/CREB	S00534	-	516	1.138000
CREB	S00144	-	517	2.549000
ATF	S01940	-	518	3.721000
ATF	S01059	-	518	1.591000



NCBI

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NEW **15 Nov 2004** Download the BLAST poster from SC2004!

About BLAST

- News
- Mailing list
- References
- NCBI Contributors

BLAST Services

- FAQs
- Program selection guide
- Web service interface

BLAST Software

- Databases
- Documentation
- Errata
- Executables
- Source code

Support

- Contact us

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

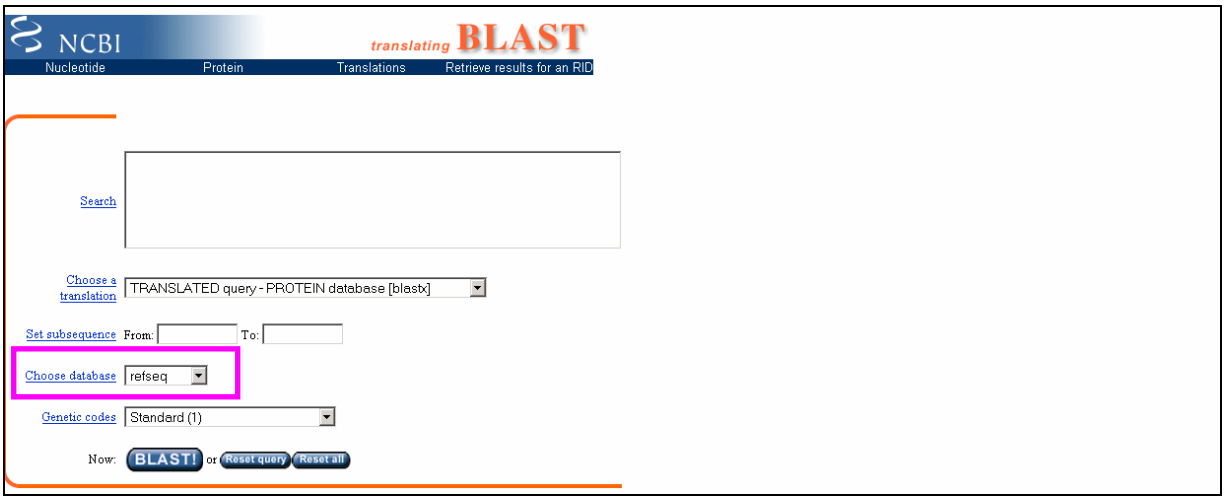
- Chicken, cow, pig, dog, sheep, cat
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

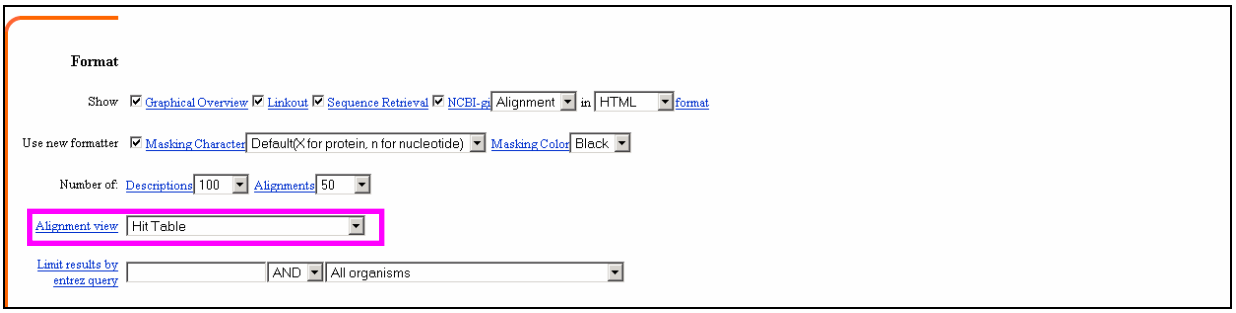
Special

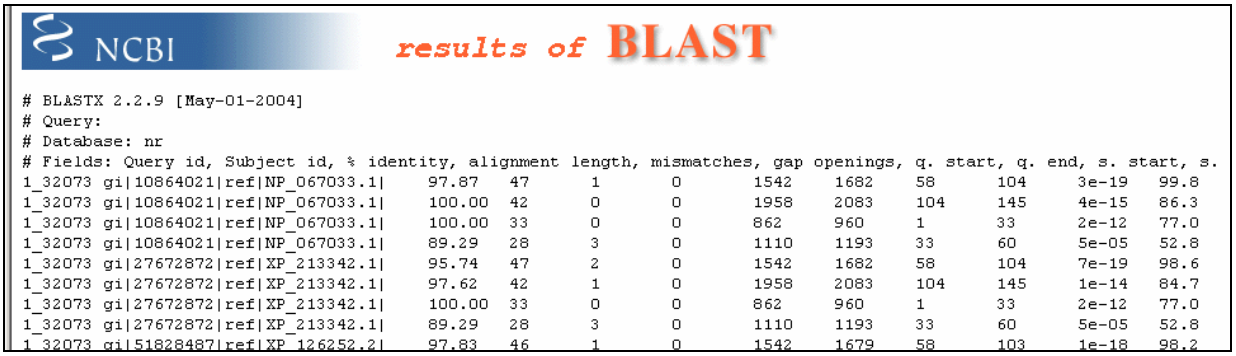
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST NEW

Meta

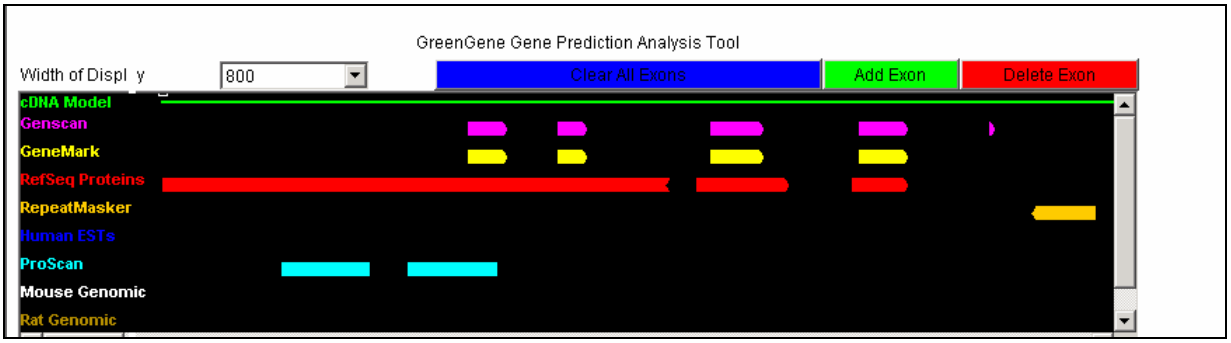
- Retrieve results by RID

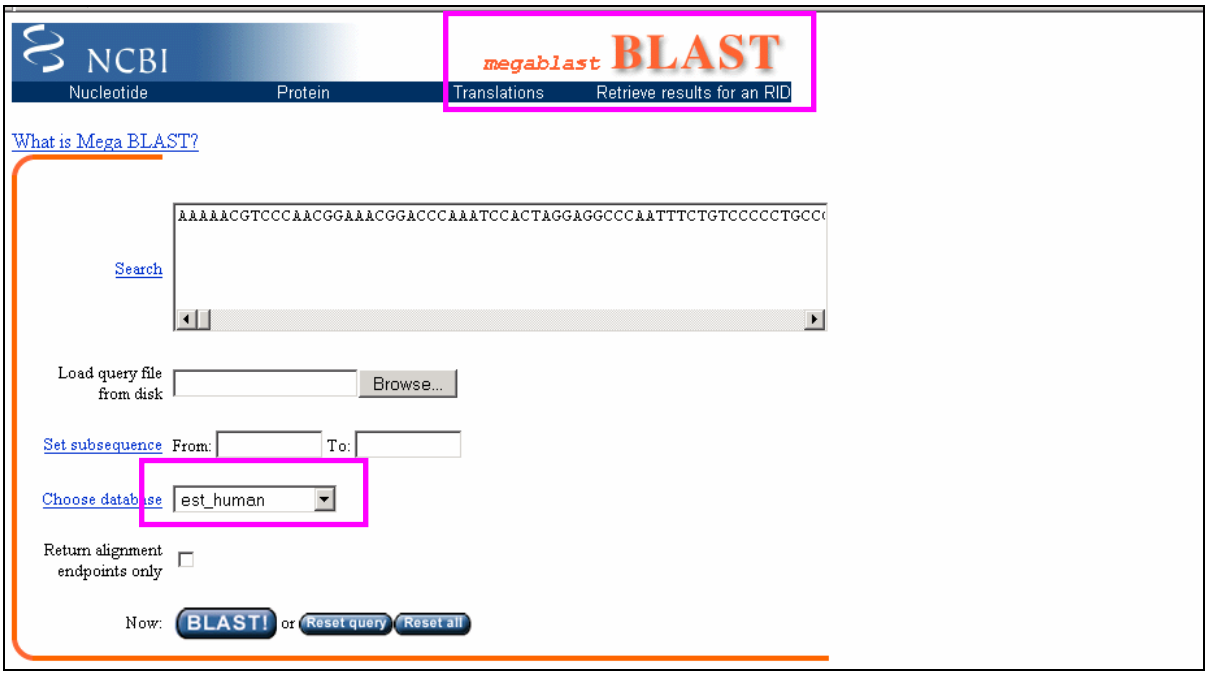


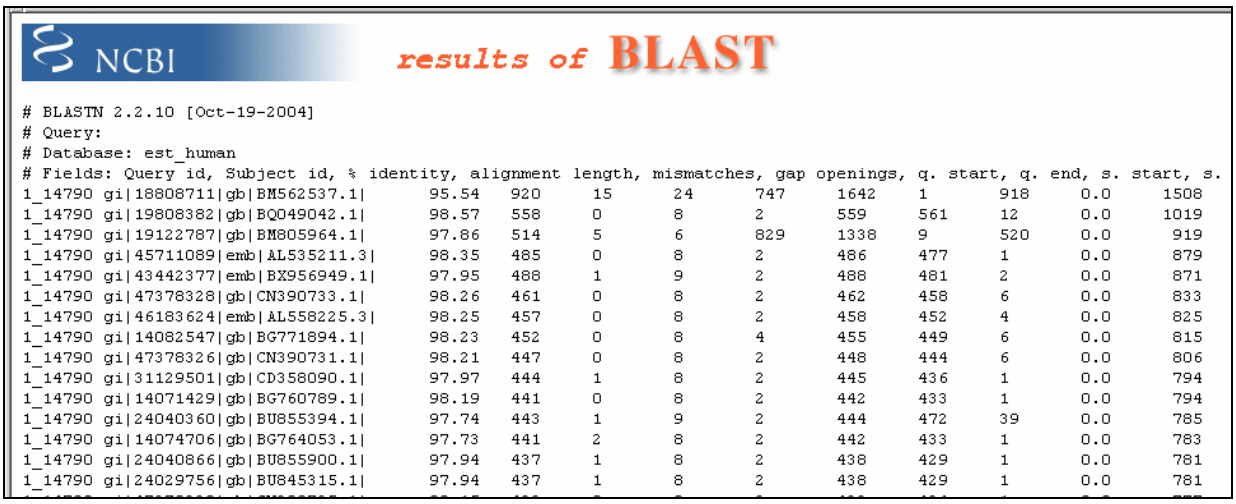




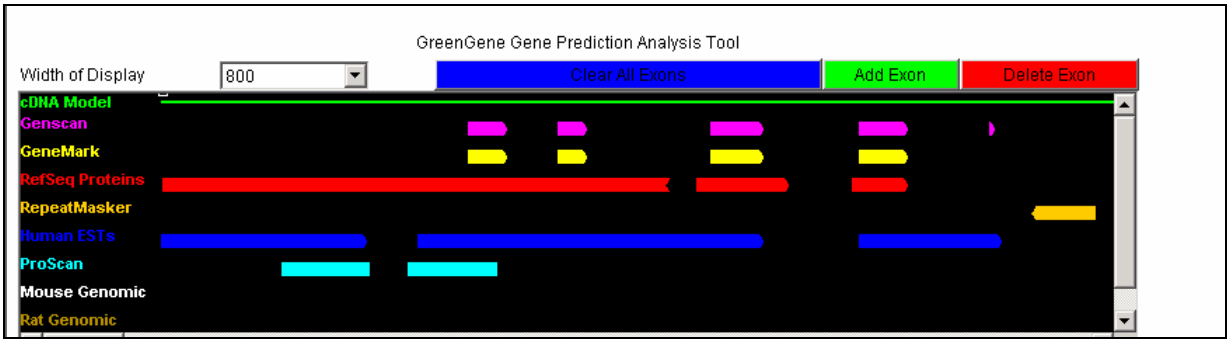

```
# BLASTX 2.2.9 [May-01-2004]
# Query:
# Database: nr
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s.
1_32073 gi|10864021|ref|NP_067033.1|    97.87   47      1       0       1542    1682    58      104     3e-19   99.8
1_32073 gi|10864021|ref|NP_067033.1|    100.00  42      0       0       1958    2083    104     145     4e-15   86.3
1_32073 gi|10864021|ref|NP_067033.1|    100.00  33      0       0       862     960     1       33      2e-12   77.0
1_32073 gi|10864021|ref|NP_067033.1|    89.29   28      3       0       1110    1193    33      60      5e-05   52.8
1_32073 gi|27672872|ref|XP_213342.1|    95.74   47      2       0       1542    1682    58      104     7e-19   98.6
1_32073 gi|27672872|ref|XP_213342.1|    97.62   42      1       0       1958    2083    104     145     1e-14   84.7
1_32073 gi|27672872|ref|XP_213342.1|    100.00  33      0       0       862     960     1       33      2e-12   77.0
1_32073 gi|27672872|ref|XP_213342.1|    89.29   28      3       0       1110    1193    33      60      5e-05   52.8
1_32073 gi|51828487|ref|XP_126252.2|    97.83   46      1       0       1542    1679    58      103     1e-18   98.2
```








```
# BLASTN 2.2.10 [Oct-19-2004]
# Query:
# Database: est_human
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s.
1_14790 gi|18808711|gb|BM562537.1|      95.54   920     15      24      747     1642    1       918     0.0     1508
1_14790 gi|19808382|gb|BQ049042.1|      98.57   558     0       8       2       559     561     12      0.0     1019
1_14790 gi|19122787|gb|BM805964.1|      97.86   514     5       6       829     1338    9       520     0.0     919
1_14790 gi|45711089|emb|AL535211.3|     98.35   485     0       8       2       486     477     1       0.0     879
1_14790 gi|43442377|emb|BX956949.1|     97.95   488     1       9       2       488     481     2       0.0     871
1_14790 gi|47378328|gb|CN390733.1|      98.26   461     0       8       2       462     458     6       0.0     833
1_14790 gi|46183624|emb|AL558225.3|     98.25   457     0       8       2       458     452     4       0.0     825
1_14790 gi|14082547|gb|BG771894.1|      98.23   452     0       8       4       455     449     6       0.0     815
1_14790 gi|47378326|gb|CN390731.1|      98.21   447     0       8       2       448     444     6       0.0     806
1_14790 gi|31129501|gb|CD358090.1|      97.97   444     1       8       2       445     436     1       0.0     794
1_14790 gi|14071429|gb|BG760789.1|      98.19   441     0       8       2       442     433     1       0.0     794
1_14790 gi|24040360|gb|BU855394.1|      97.74   443     1       9       2       444     472     39      0.0     785
1_14790 gi|14074706|gb|BG764053.1|      97.73   441     2       8       2       442     433     1       0.0     783
1_14790 gi|24040866|gb|BU855900.1|      97.94   437     1       8       2       438     429     1       0.0     781
1_14790 gi|24029756|gb|BU845315.1|      97.94   437     1       8       2       438     429     1       0.0     781
```



NCBI

megablast BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

What is discontiguous Mega BLAST?

Search: GGACCATCTTCCCAATTCCGTTCAGAAAATATTCAAGCTGCAAACCCTTCAGCAGCCGGGGC

Load query file from disk: Browse...

Set subsequence: From: To:

Choose database: chromosome

Return alignment endpoints only

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query: AND Mus musculus [ORGN]

GreenGene Gene Prediction Analysis Tool

Width of Display: 800 | Clear All Exons | Add Exon | Delete Exon

Genscan
GeneMark
RefSeq Proteins
RepeatMasker
Human ESTs
ProScan
Mouse Genomic
Rat Genomic

NCBI

megablast BLAST

Nucleotide Protein Translations Retrieve results for an RID

What is discontiguous Mega BLAST?

Search AAAAACGTCCCAACGGAAACGGACCCAAATCCACTAGGAGGCCCAATTTCTGTCCCCCTGCC

Load query file from disk Browse...

Set subsequence From: To:

Choose database chromosome

Return alignment endpoints only

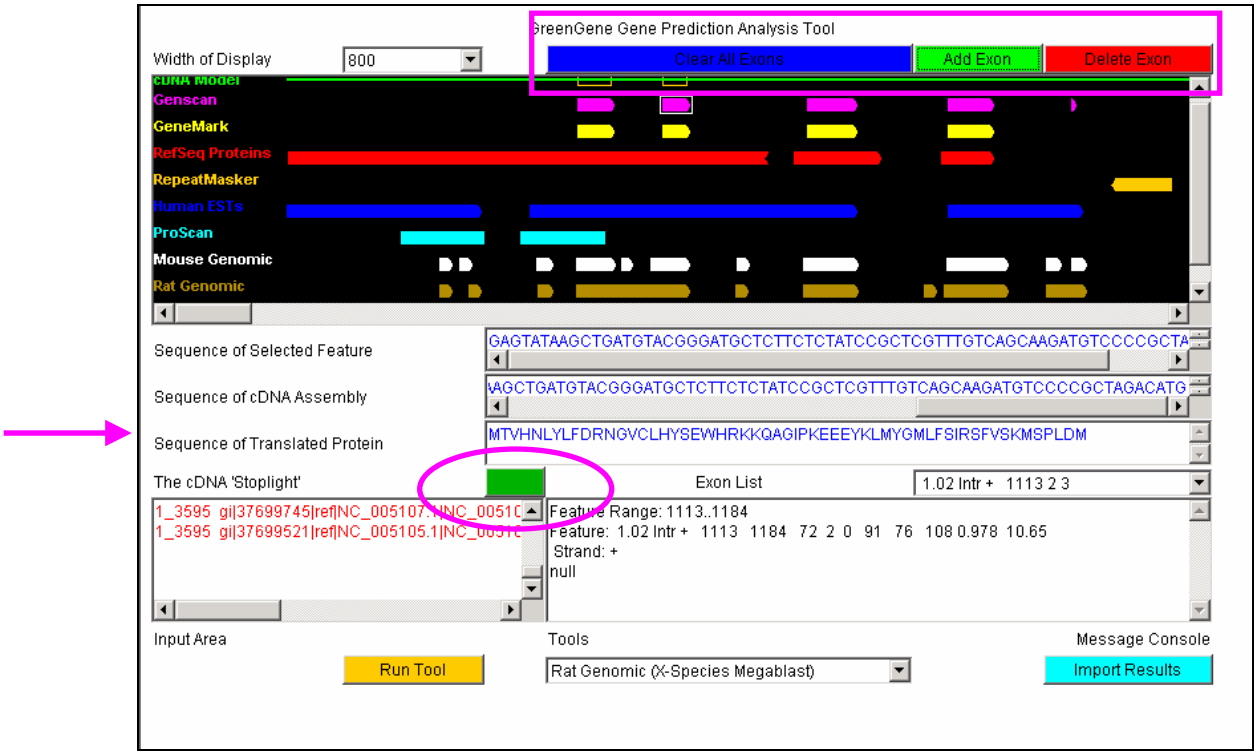
Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query AND Rattus norvegicus [ORGN]

GreenGene Gene Prediction Analysis Tool

Width of Display 800 Clear All Exons Add Exon Delete Exon

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

Create a custom coding sequence and the corresponding protein sequence: