

## APPENDIX 2 - BIOINFORMATICS (PARTS I AND II)

**HC70AL Spring 2004**

**An Introduction to Bioinformatics -- Part I**

**By**

**Brandon Le**

**April 6, 2004**

### **What are the Characteristics of a Gene?**

- **An ordered sequence of nucleotides**
- **A unique position/location in the genome**
- **Polarity (5' to 3')**
- **Exons and Introns**

## What are the Anatomical Features of Genes?

- Discrete beginning and discrete end
- Two strands of DNA
- Double helical
- Strand one (5' to 3')
- Strand two (3' to 5')
- Sense strand (5' to 3')
  - specifies the trait
- Nonsense strand (3' to 5')
  - template for transcription

### Sense Strand

```
5' - ACGTCAGTCGATGCATGCTAGCTAGC - 3'  
3' - TGCAGTCAGCTACGTACGATCGATCG - 5'
```

### Nonsense Strand

## Genes Have a Unique Position in the Genome!

**Task:** Where is your gene located in the genome?

**Tools:** The Arabidopsis Information Resources (TAIR)  
(<http://www.arabidopsis.org>)

### Procedure:

1. Select Seqviewer
2. Enter gene number (ex. AT1G18260)
3. Submit

### Results/Question:

1. What chromosome is your gene in?
2. What other genes/markers are next to your gene?
3. What is the exact position of your gene in the genome?

01 AT2G22800  
02 AT2G23290  
03 AT2G37120  
04 AT3G09735  
05 AT3G12840  
06 AT3G50060  
07 AT3G53370  
08 AT4G37260  
09 AT4G37790  
10 AT5G03220  
11 AT5G03500  
12 AT5G19490  
13 AT5G67300

## **Genes Have a Unique Order of Nucleotides!**

**Task:** What is the order of nucleotides for your gene?

**Tools:** The Arabidopsis Information Resources (TAIR)  
(<http://www.arabidopsis.org>)

**Procedure:** (Continue from previous slide)

1. Click on Location

**Results/Question:**

1. What are your neighbor genes?
2. What is the orientation of your gene?
3. How big is your gene?

## **Genes Have Exons and Introns!**

**Task:** How many exons and introns does your gene have?

**Tools:** The Arabidopsis Information Resources (TAIR)  
(<http://www.arabidopsis.org>)

**Procedure:** (Continue from previous slide)

1. Click on gene information on the right

**Results/Question:**

1. How many exons/introns in your gene?
2. What are exons?
3. What are introns?

## Gene Encodes a Protein

**Task:** Determine the protein encoded by gene?

**Tools:** The Arabidopsis Information Resources (TAIR)  
(<http://www.arabidopsis.org>)

**Results/Question:**

1. How large is your protein?
2. What are the anatomy of a protein?

N-terminal



C-terminal

## What is the identity of your gene?

**Task:** What does your gene code for?

**Tools:** NCBI BLAST Tools  
(<http://www.ncbi.nlm.nih.gov/BLAST>)

## What is BLAST?

### Basic Local Alignment Search Tool (BLAST)

## What does BLAST do?

A family of programs that allows you to input a query sequence and compare it to DNA or protein sequences in db.

## What are the steps to performing BLAST search?

Paste sequence of interest into BLAST input box

Select BLAST program

Select db

Select Optional Parameters

The screenshot shows the NCBI translating BLAST search interface. At the top, there is a navigation bar with the NCBI logo and the text "translating BLAST". Below the navigation bar, there are four tabs: "Nucleotide", "Protein", "Translations", and "Retrieve results for an RID". The "Protein" tab is selected. Below the tabs, there is a search box with a "Search" button to its left. Below the search box, there are several options: "Choose a translation" with a dropdown menu showing "TRANSLATED query - PROTEIN database [blastx]", "Set subsequence" with "From:" and "To:" input fields, "Choose database" with a dropdown menu showing "nr", and "Genetic codes" with a dropdown menu showing "Standard (1)". At the bottom, there are three buttons: "BLAST!", "Reset query", and "Reset all".

## What are the different BLAST Programs?

Fastest

**blastp** - protein query vs protein db

**blastn** - DNA query vs DNA db

**blastx** - translated DNA query vs protein db

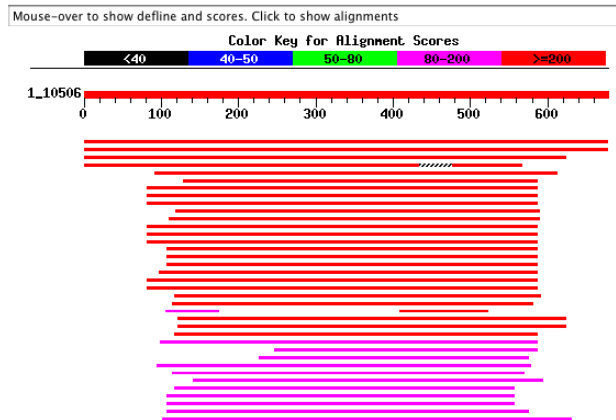
**tblastx** - protein query vs translated DNA db

Slowest

**tblastn** - translated DNA query vs translated DNA db

## Anatomy of a BLAST Result -- Part I

### Distribution of 339 Blast Hits on the Query Sequence



## Anatomy of a BLAST Result -- Part II

Sequences producing significant alignments:	(bits)	Value
<a href="#">gi 14532716 gb AAK64159.1</a> unknown protein [Arabidopsis tha...	1206	0.0
<a href="#">gi 18394588 ref NP_564049.1</a> suppressor of lin-12-like prot...	1209	0.0
<a href="#">gi 15219499 ref NP_177498.1</a> suppressor of lin-12-like prot...	877	0.0
<a href="#">gi 11120786 gb AAG30966.1</a> hypothetical protein, 3' partial...	426	e-118
<a href="#">gi 41151276 ref XP_046437.5</a> chromosome 20 open reading fra...	291	3e-77 <a href="#">L</a>
<a href="#">gi 13559241 emb CAB65792.2</a> dJ842G6.2 (novel protein imilar...	282	2e-74 <a href="#">L</a>
<a href="#">gi 19923669 ref NP_005056.3</a> sel-1 suppressor of lin-12-lik...	268	4e-70 <a href="#">L</a>
<a href="#">gi 6851089 gb AAF29413.1</a> SEL1L [Homo sapiens] >gi 17646138...	268	4e-70 <a href="#">L</a>
<a href="#">gi 9967440 dbj BAB12403.1</a> SEL1L [Mesocricetus auratus]	264	4e-69
<a href="#">gi 31203035 ref XP_310466.1</a> ENSANGP00000019196 [Anopheles ...	263	1e-68
<a href="#">gi 21355295 ref NP_651179.1</a> CG10221-PA [Drosophila melanog...	263	1e-68 <a href="#">L</a>
<a href="#">gi 20857527 ref XP_127076.1</a> Sell (suppressor of lin-12) 1 ...	261	4e-68 <a href="#">L</a>
<a href="#">gi 4159995 gb AAD05210.1</a> SEL1L [Mus musculus] >gi 20073079...	259	1e-67 <a href="#">L</a>
<a href="#">gi 29336095 ref NP_808794.1</a> Sell (suppressor of lin-12) 1 ...	259	2e-67 <a href="#">L</a>
<a href="#">gi 29612522 gb AAH49959.1</a> Sellh protein [Mus musculus]	258	4e-67 <a href="#">L</a>
<a href="#">gi 17563256 ref NP_506144.1</a> Suppressor/Enhancer of Lin-12 ...	247	9e-64 <a href="#">L</a>
<a href="#">gi 1255199 gb AAC47112.1</a> sel-1 gene product	247	9e-64 <a href="#">L</a>

## Anatomy of a BLAST Result -- Part III

```

>gi|14532716|gb|AAK64159.1 unknown protein [Arabidopsis thaliana]
Length = 678

Score = 1206 bits (3120), Expect = 0.0
Identities = 614/678 (90%), Positives = 614/678 (90%)

Query: 1 MRILSYGIVILSLLVFSFIEFGVHARPVVLVXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 60
MRILSYGIVILSLLVFSFIEFGVHARPVVLV V
Sbjct: 1 MRILSYGIVILSLLVFSFIEFGVHARPVVLVLSNDDLNSGGDDNGVGESSDFDFGESEP 60

Query: 61 XXXXXLDPGSWRSIFEPDDSTVQAASPOYYSGLKKILSAASEGNFRLMEEAVDEIEAASS 120
LDPGSWRSIFEPDDSTVQAASPOYYSGLKKILSAASEGNFRLMEEAVDEIEAASS
Sbjct: 61 KSEELDPGSWRSIFEPDDSTVQAASPOYYSGLKKILSAASEGNFRLMEEAVDEIEAASS 120

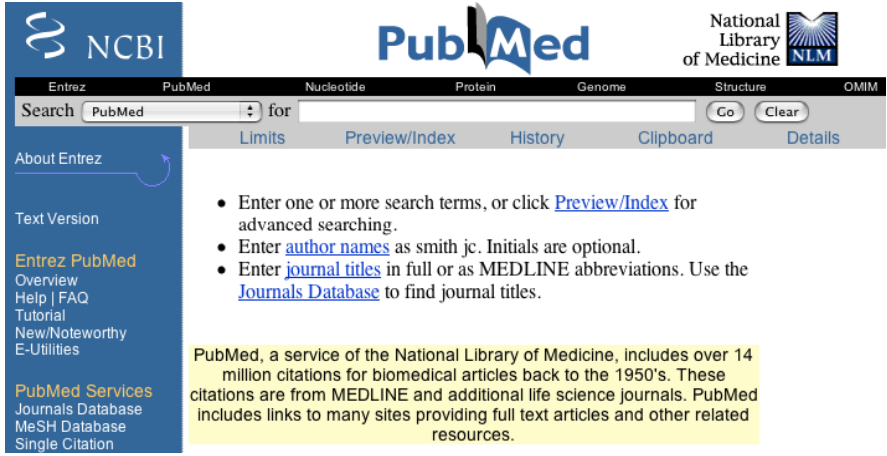
Query: 121 AGDPHAQSIMGFVYIGIMMREKS KSKSFLHNNFAAAGNMQSKMALAFTYLRQDMHDKAV 180
AGDPHAQSIMGFVYIGIMMREKS KSKSFLHNNFAAAGNMQSKMALAFTYLRQDMHDKAV
Sbjct: 121 AGDPHAQSIMGFVYIGIMMREKS KSKSFLHNNFAAAGNMQSKMALAFTYLRQDMHDKAV 180

Query: 181 QLYAELAETAVNSFLISKDSPVVEPTRIHSGTEENKALRKS RGEEDDFQILEYQAQKG 240
QLYAELAETAVNSFLISKDSPVVEPTRIHSGTEENKALRKS RGEEDDFQILEYQAQKG
Sbjct: 181 QLYAELAETAVNSFLISKDSPVVEPTRIHSGTEENKALRKS RGEEDDFQILEYQAQKG 240

Query: 241 NANAMYKIGLFYFGLRGLRRDHTKALHWFLKAVDKGEPRSMELLGEIYARGAGVERNYT 300
NANAMYK GLFYFGLRGLRRDHTKALHWFLKAVDKGEPRSMELLGEIYARGAGVERNYT
Sbjct: 241 NANAMYKNGLFYFGLRGLRRDHTKALHWFLKAVDKGEPRSMELLGEIYARGAGVERNYT 300

```

## PubMed - Endless Resources



The screenshot displays the PubMed website interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos is a navigation bar with tabs for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, and OMIM. The search bar contains the text "PubMed" and "for", with "Go" and "Clear" buttons. Below the search bar are links for Limits, Preview/Index, History, Clipboard, and Details. On the left side, there is a sidebar with links for About Entrez, Text Version, Entrez PubMed (Overview, Help | FAQ, Tutorial, New/Noteworthy, E-Utilities), and PubMed Services (Journals Database, MeSH Database, Single Citation). In the center, there are three bullet points providing search instructions: "Enter one or more search terms, or click [Preview/Index](#) for advanced searching.", "Enter [author names](#) as smith jc. Initials are optional.", and "Enter [journal titles](#) in full or as MEDLINE abbreviations. Use the [Journals Database](#) to find journal titles." Below these instructions, a yellow highlighted box contains the text: "PubMed, a service of the National Library of Medicine, includes over 14 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources."



**HC70AL Spring 2004**

**An Introduction to Bioinformatics -- Part II**

**By**

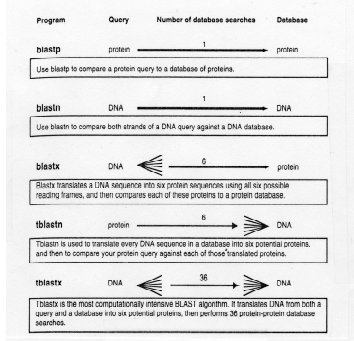
**Brandon Le**

**April 8, 2004**

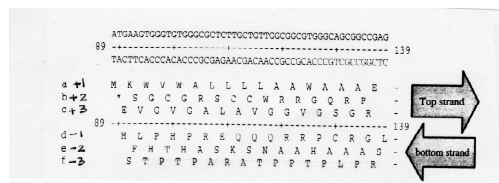
**Review of BLAST Search**

1. **What is the purpose of running BLAST Search?**
2. **What are the steps to performing BLAST search?**
3. **What does the e-value from a blast result tell you?**
4. **How may BLAST program can you perform?**
5. **What BLAST program(s) takes the least computational time?**
6. **What BLAST program(s) takes the most computational time? Why?**

## What are the Five BLAST Search Programs?



•How many proteins can a short DNA sequence potentially encode?



## Question:

You have DNA Sequence. You want to know which protein in the main protein database is most similar to some protein encoded by your DNA.

Which BLAST program should you use?

Suppose you have a protein sequence. Which BLAST program should you use?

## HOW to interpret BLAST results?

Identifier of protein in database

Expect =  $8e-79$   
Probability of finding this alignment in a database by chance and chance alone.

Origin of protein sequence

What does frame mean?

Exact match of amino acid

What is query?  
what is subject?

What does the "+" mean?

What does this number correspond to?

positively charged amino acids

K = Lysine  
R = Arginine

(4)

```
gi|1355241|mml|CAB5732.2| d|8422d.2 (novel protein similar to REX12 (swi-1  
ITN-12, C. elegans)-like) (nomo saplenn) -----  
Covrlnh = 289 # of amino acid in the protein  
Identifiers = 177/468 (37%) Positively = 269/468 (57%) Gaps = 9/468 (1%)  
Frame = +1  
Query: 388 HGFV-YGIMHREKREKFLHHPFAAGDGGHSHALAPFTLRQ---DHDRKAVULT 589  
+GF+ YIGM E ++K+ +F+ K+AGNH S+N L T IL -----AY+ Y  
Sbjct: 2 HFLSDYDIDH--YDGAALITTTGAGGHHHGGHITGVITLGLSIVLGDVRSVLSV 59  
Query: 658 ARLLAFAWQSPFTKEDGQVDFPIHRRPESNGALRKRREDEDFDLEKQAGQGAN 729  
+AA+ ++F S+ PV E R+ --E+ + + +G ++A+G+  
Sbjct: 60 KRVDYADTFERSQVFPV-EKVL---TERPELSSHELDWDIYQYKFLAERGVD 115  
Query: 730 ANKIGLFTYFGLRLRDRFKALHFLKAVYDKEFSEHLLRISFANQ-AGEVDF 905  
+G+ +G+GL +D+ KALHYEKA G +R +D++ G A V +N A) 905  
Sbjct: 116 IQVHGGHLIIGHMELDQYFKALDYFKLAAAGADHAPVGHYFAGHAFVGHMA 175  
Query: 967 IHWFLAKRSELKSPKPTIETLVEVEVDKRFVARRVFKAUNEDSDGHLML 1085  
++AA +G +G+ LY E GV NY +A +E+KA + P + LG +Y  
Sbjct: 176 PKYFMAASKGNAIGLGLLTFHGRVFL-NTAALKTFQAAEKQFPDQQLGPHY 234  
Query: 1087 LKCIQVDFGQATFTFFVAHQGKPFYGLAKHTYQVLEHLEHATTFELVAEHL 1285  
G G+ +D+ A RFF++T+ +GK A Y LAKN+ TG G+ ++ A --K V E G  
Sbjct: 235 YGSHIHWYKALAFKTFYVLSQGGPLATYLLAHVATGQVYVGHGTVLSEDFV 294  
Query: 1567 DWETRWALRFLKINDVIALIVFNARMEVVAQSHMAWILDEYGEREHCQVSRPC 1446  
W+ A Y GD+ +L+ Y+ +AENGVAQSN+IL+ -----) 1446  
Sbjct: 295 HWAEKFLYATYATKGGIDDSLVQYALLAENGVEVAQSNAP ILESKANIL(-----) 346  
Query: 1447 TDKER-HERASLWASDQNEHALLIGDATTYQKTERDFRAGAY-HRARSQNA 1620  
+ES+ + A LW RAY QUN A + IGD+ TGG GT++D+ AA Y + A HA  
Sbjct: 347 -ERKHYFALLLWRAAQGNAPARVIGDHYVQGTNDGQTATYVZAAHVVHA 405  
Query: 1471 QAHFTEVDRSRRQDFDFRHLGKLVYRHLRQDAARFQTLALGL 1766  
QAHFL YH+ESG G+ D+HL+H YD + Q+ A +PV A+ L 1766  
Sbjct: 406 QAHFLAYHYEGLGTTKIDHLGLYDHAAGTSPDRIPLVFAVRL 453
```

## Review of gene transcription

1. What product is made after transcription?
2. How is the product similar/different from the gene?
3. What is cDNA?
4. What important information does a cDNA tell you about a gene?
5. What are ESTs?
6. What important information does ESTs tell you about a gene?

## Annotation of your gene

1. What chromosome is your gene in?
2. How “big” is your gene?
3. How many exons and introns in your gene?
4. What orientation is your gene in the genome?
5. What is the specific position of your gene in the genome?
6. What gene is “upstream” of your gene?
7. What gene is “downstream” of your gene?
8. How far are the other genes (6 & 7) from your gene?
9. What is the “structure” of your gene?
10. What is the size of the protein in your gene encodes?
11. What protein does your gene encode
12. Is your gene structure predicted by a program?

## Webbook - A Virtual Lab Notebook

Webbook is a web lab notebook

**Purpose/goal:** To have access to experiments carried out by  
Lab members, etc... from anywhere  
Also serves as a repository for protocols, stocks/reagents

**Created by:** Harry Hahn  
Brandon Le  
Bob Goldberg

<http://estdb.biology.ucla.edu/webbook>

## Using the Webbook

- 1. Username: email username**  
**Password: 9 digit student id**
- 2. Check message board for important news/updates**
- 3. An overview of the different sections**
  - Projects** - list of experiments
  - Stocks** - catalog of stocks/reagent in the lab
  - Protocols** - procedures carried out in the lab (pdf format)
  - Calendar** - calendar to plant your experiments
  - Browse** - search and look at other members experiments
  - Contact** - email for help
  - Logout** - will logout if idle for 30 min

## Webbook Login Page

webBOOK

Help Login

webBOOK Login

Username: ble Password: Login

*Last modified August 03 2003 21:16:09.*  
*Copyrighted by the University of California (2003)*  
*Created by Harry Hahn and Brandon Le, Laboratory of Bob Goldberg, UCLA*

## Creating Projects / Experiments

1. Title of project
2. Questions/Purpose of project
3. Summary of project (ideas)

## Entering Gene Information

Genes

---

Create gene

Fields marked with a red asterisk (\*) are **REQUIRED**

Gene Name:*	<input type="text"/>
Species:	<input type="text"/>
Sequence:	<input type="text"/>
Sequence Type:*	-- Select --
Amino Acid Sequence:	<input type="text"/>
Chromosome:	<input type="text"/>
EST Data:	<input type="text"/>
Functional Category:	<input type="text"/>
Promoter:	<input type="text"/>
Domains:	<input type="text"/>
Hits:	<input type="text"/>
Attach a file:	Title: <input type="text"/>
	File: <input type="button" value="Choose File"/> no file selected
	Description: <input type="text"/>

## Entering Experiments Information Part 1

Experiments	
Fields marked with a red asterisk (*) are <b>REQUIRED</b>	
<b>Title:*</b>	<input type="text"/>
<b>Goal:*</b>	<input type="text"/>
<b>Background Info:*</b>	<input type="text"/>
<b>Approach:*</b>	<input type="text"/>
<b>Controls:*</b>	<input type="text"/>
<b>Discussion:</b>	<input type="text"/>
<b>Next:</b>	<input type="text"/>

## Entering Experiment Information Part II

<b>Materials</b>	<b>Primer *</b> <input type="text" value="AT2G22800-FW&lt;br/&gt;AT2G22800-RV&lt;br/&gt;AT2G23290-FW&lt;br/&gt;AT2G23290-RV&lt;br/&gt;AT2G37120-FW&lt;br/&gt;AT2G37120-RV&lt;br/&gt;AT3G09735-FW&lt;br/&gt;AT3G09735-RV"/>
<b>Protocols:</b>	<b>Protocols *</b> <input type="text" value="*Sequencing Using SPPCR&lt;br/&gt;Alkali Lysis Plasmid Isolation&lt;br/&gt;Arabidopsis Tissue Harvest for GeneChIP Experiment&lt;br/&gt;Bacteria Chromosome Mini-Prep&lt;br/&gt;Bacteriophage&lt;br/&gt;Chromatin Immunoprecipitation with Leaves from Arabidopsis"/>
<b>Attach a file:</b>	<b>Title:</b> <input type="text"/> <b>File:</b> <input type="button" value="Choose File"/> no file selected <b>Description:</b> <input type="text"/> <small>All files must have a file name extension. Images must end in .jpg, .png, or .gif. Additional files can be attached by later editing this record.</small>

## Entering References Relating to your Gene

### References

#### Create reference record

Fields marked with a red asterisk (\*) are **REQUIRED**

<b>Author(s):*</b>	<input type="text"/>
<b>Title:*</b>	<input type="text"/>
<b>Journal:*</b>	<input type="text"/>
<b>Year*</b>	<input type="text"/>
<b>PDF File</b>	<input type="button" value="Choose File"/> no file selected

---