Application for Automating Database Storage of EST to Blast Results

Vikas Sharma
Shrividya Shivkumar
Nathan Helmick
Outline

- Biology Primer
  - Vikas Sharma
- System Overview
  - Nathan Helmick
- Creating ESTs
  - Nathan Helmick
- Automating Blast Search of ESTs
  - Shrividya Shivkumar
- Import Blast Results to Database
  - Vikas Sharma
- Issues – Enhancements – Deliverables
  - Nathan Helmick
Biology Primer

Vikas Sharma
What are EST

- They represent a snapshot of genes expressed in a given tissue and/or at a given developmental stage.

- They are tags of expression for a given cDNA library.

- That’s why they are known as Expressed sequenced tags
ESTs: GENE DISCOVERY MADE EASIER

► ESTs provide quick and inexpensive route for discovering new genes.

► For obtaining data on gene expression and regulation.

► For constructing genome maps.

✓ In 1992, first EST data appeared in GenBank.

✓ As of July 2008, dbEST contained more than 54 million sequences, which is 62 % of all GenBank entries.
How Are They Made?

► ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long).

► Generated by sequencing either one or both ends of an expressed gene.

► The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "tags" to fish a gene out of a portion of chromosomal DNA by matching base pairs.
Central Dogma

![Diagram of the Central Dogma](image)
Formation of a cDNA Library

1. Isolate and collect mRNA
2. Reverse transcriptase
3. RNA → DNA
4. Insert into bacterial plasmids
5. Insert plasmids into bacteria
6. Grow
7. Isolate plasmid and purify DNA
8. DNA
9. Sequence

ATCCTATCGTAGCTGGCCTGACTTTA
Chemistry behind sequencing

Primer for replication

Strand to be sequenced

Prepare four reaction mixtures; include in each a different replication-stopping nucleotide

C  G  A  T

Primed DNA

Replication products of “C” reaction

Separate products by gel electrophoresis

Read sequence as complement of bands containing labeled strands
ABI 377 DNA Sequencers
FUNCTIONS AND CAPABILITIES of the ABI 377 DNA SEQUENCER:

✓ 900 bases with 98.5% accuracy
✓ 24–96 samples throughput
✓ Run sequencing reactions based on dye-labeled terminator chemistry
✓ Reads up to and beyond 900 bases per sample generating long reads
Disadvantage of EST

- Data produced are of not high quality. Due to substitutions, deletions, or insertions in EST sequences compared with the parent mRNA sequence.

- EST region only between the 100 and 300 sequence may be the most accurate part of the sequence.

- There is a need for removal of vector sequences present in EST sequences.
Figure 1. Distribution of average contamination ratio of sequencing centers. Sequencing centers were classified according to the number of their total sequences in dbEST and were calculated an average contamination ratio of each class. The x-axis represents the classes of sequencing centers and y-axis represents their contamination ratio. The average contamination ratio is lower for centers that have submitted a larger number of sequences. Small sequencing centers (<10,000 ESTs) have more than double the contamination of large sequencing centers (>1,000,000 ESTs).

http://nar.oxfordjournals.org/cgi/content/full/gkn648
System Overview

Nathan Helmick
System Flow Overview

Develop ESTs
- Covert input data from phd to fasta format
  - Import reads into database
- Remove sequencing vector from reads
- Generate EST through PHRAP
  - Import EST into database

Run Blast Search
- Run Blast search for each EST
  - Store results to temporary text file

Load Blast Results to Database
- Parse blast search results file and import into database
Database Design

- The guiding principle of design for the database was to maintain as much data as possible.
- Some of our data may be of little/no interest to our end users – potentially could be revised for smaller database size.
- FASTA format files contain multiple reads per file and a separate FASTA format file with quality values.
Converting Inputs To ESTs

Nathan Helmick
Conversion PHD to FASTA

- PHD and FASTA are both standards for storing sequence and quality information.

- PHD files contain a single read with three sections: header information, sequence data, and quality.

- FASTA format files contain multiple reads per file and a separate FASTA format file with quality values.

Cloning Vectors

► A vector is an agent (plasmid, yeast artificial chromosome, others) capable of injecting a DNA fragment into a host

► If such agent is used to reproduce the vector, it is known as a cloning vector

► These cloning vectors are frequently used by biologist to reduce the need for large quantities of DNA material

Why Remove Cloning Vectors

► If cloning vectors are not removed from the data set, they can corrupt the final ESTs

► Cloning vectors can complicate the process of reassembly

► Cloning vectors can cause identical reads to look different

Cross Match

► Cross Match is the tool that we have selected for removing cloning vectors

► It accepts a vector file containing the sequence used for cloning
  ▪ Many labs only create a single sequence file that contains all potential vectors they may use for cloning

► If segments of a sequence are determined to be from the cloning vector a series ‘x’s are placed in the effected region

Sequences We Receive

► The sequences we receive are reads, read by Applied Biosystems 3730 DNA Analyzer

► The consist of mRNA reads taken from targeted organisms

► The maximum size of each read depends on speed settings, but is a maximum of 900 base-pairs
Glossary of Genetics, Interaction of mRNA in cell, Licensed under creative commons
Why read mRNA?

► We could just analyze the DNA
  ▪ Would give us more complete picture of the organism

► Algorithms for determining ESTs from DNA sequences have not advanced to the point that they can sufficiently predict expressed genes

► Essentially analyzing the mRNA allows us to let the organism do the first pass filtering of the DNA sequence – it will tell us exactly what genes are being expressed
PHRAP

- PHRAP (Phil’s Revised Assembly Program) is a program designed primarily for the reassembly of shotgun sequenced DNA.

- It is an industry standard and provides very good sequencing when accurate quality values are available.

- Maybe freely used (without support) for non-commercial ventures.
Dealing with Large Input (PHRAP)

- PHRAP uses a modified Smith-Waterman algorithm
- Running Smith-Waterman on the entire sequence would be prohibitively time consuming
- PHRAP runs Smith-Waterman only in banded areas that meet a minimum exact match count

Why Use PHRAP?

► Even though we aren’t exactly reading shot-gun blasted DNA, it is possible our genes could extend the capacity of our analysis equipment
  - Current data set does not appear to be the case

► PHRAP will identify and combine repeated reads of the same gene or segments of the same gene

► Currently we get about 1 contig (combination or reads) for each singlet (single read) in our final EST set
Analysis of PHRAP vs Others

► It is difficult to compare PHRAP with others as it relates to its ability to analyze EST sequences
  ▪ Many other queries develop their own qualitative values
  ▪ Much of the research comparing sequencing software is done using data from GeneBank or equivalent databases, which often doesn’t contain quality data

► Some competitors to PHRAP include CAP3, TA-EST, and TIGR Assembler

► PHRAP is more sensitive than other algorithms to detecting slight differences in genes
  ▪ PHRAP only considers reads identical if 95% exact
  ▪ Most other algorithms are in the high 60% - low 70%

Analysis of PHRAP vs Others

Why Perl for Automation

► Perl is very simple

► Perl provides many built-in libraries for interfacing to databases

► Perl is widely used in the biological communities

► Perl provides very strong support for Regular Expressions
#!/usr/bin/perl –w

while( (<SEQUENCE_FILE>) ) {
    # Remove the new line character if it is there
    chomp;
    if( />/ ) {
        insertRead( $fileName, $chem, $dye, $time, $sequence );
    }
}

@line = split / /, _;
$fileName = $line[4];
$chem = $line[6];
$dye = $line[8];
$time = join " ", $line[11], $line[12], $line[14], $line[13];
$sequence = "";
}
#!/usr/bin/perl –w

# This loop reads each line of the file into the default variable ($_).
while( (<SEQUENCE_FILE>) ) {
    # Remove the new line character if it is there
    chomp;

    # Check if this is the start of a new read – the comparison is done with the
    # default variable ($_) since nothing else is defined
    if( />/ ) {
        insertRead( $fileName, $chem, $dye, $time, $sequence );
    }
}

# Split the info line into an array of tokens separated by space, then
# load the specific variables with the correct token
@line = split / /, $_;
$fileName = $line[4];
$chem = $line[6];
$dye = $line[8];
$time = join " ", $line[11], $line[12], $line[14], $line[13];
$sequence = "";
}
Performing BLAST Search On ESTs

Shrividay Shivkumar
BLAST

► Local sequence alignment tool
► Used to compare nucleotide or protein sequences.
► Helps in determining members of the gene families, evolutionary relationships.
► Different types of blast - blastp, blastn, blastx, tblastn, tblastx.
BLASTX

- Takes an input nucleotide sequence
- Translates the query sequence in 6 reading frames
- Performs a blastp search with each reading frame
- Displays output

How blastp works?

► Remove low-complexity region or sequence repeats in the query sequence.
  - Sequences with unusual composition – these can create problems in sequence similarity searching.
    - PPCDPPPPPPKDKKKKDDGPP
    - AAATAAAATAATAAAAAAT

► Replaces the repeats with
  - X – In protein sequence
  - N – In DNA sequence
How blastp works?

► Make a k-letter word list of the query sequence.

Query sequence: PQGEFG

Word 1: PQG

Word 2: QGE

Word 3: GEF

Word 4: EFG
How blastp works?

► List the possible matching words.
► Keep track of the high-scoring words.
► Repeat above steps for each 3-letter word in the query sequence.
► Scan the database sequences for exact match with the remaining high-scoring words.
How blastp works?

► Extend the exact matches to high-scoring segment pair (HSP).

```
Query sequence: R P P Q G L F
Database sequence: D P P E G V V

Score: -2 7 7 2 6 1 -1

Optimal accumulated score = 7+7+2+6+1 = 23
```
How blastp works?

► List all of the HSPs in the database whose score is high enough to be considered.
► Report the matches whose expect score is lower than a threshold parameter E.
How blastp works?

The BLAST Search Algorithm

query word (W = 3)

Query: GSVEDTTGQSALAILNCKITRPLVLNDKQLMDURIEERLNVEAFVEDAEELRQLQEDL

neighborhood words

PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PHG 13
PQA 12
PQN 12
e tc...

neighborhood score threshold (T = 13)

Query: 325 SLLAILNCKITRPLVLNDKQLMDURIEERLNVEA 365
+LA++L+ TP GR++ P+ D + ER + A
Subject: 290 TLASVLDCVTNGSRLKFWLHMVRDTKVLLEQQTIGA 330

High-scoring Segment Pair (HSP)
BLAST Statistics

Approx. 50 words are found for every residue in a protein sequence.

- So, for a sequence of length 250 the total number of words will be $250 \times 50 = 12500$

Probability of finding a sequence having score $\geq S$

where

$$1 - e^{-y}$$

$$y = Kmn e^{-\lambda S}$$

$m, n =$ length of the input sequences.

$K, \lambda =$ parameters of statistical significance.
BLAST Statistics

- Compiling the list of high-scoring words (W)
- Scanning the database for hits
- Extending the hits
Figure 2. The central processing unit time required to execute BLAST on the PIR protein database (Release 23.0) as a function of the size of the word list generated.
# Blast Statistics

Table 1

The probability of a hit at various settings of the parameters \( w \) and \( T \), and the proportion of random MSPs missed by BLAST

<table>
<thead>
<tr>
<th>( w )</th>
<th>( T )</th>
<th>Probability of a hit ( \times 10^3 )</th>
<th>Linear regression (-ln(q) = aS + b)</th>
<th>Implied % of MSPs missed by BLAST when ( S ) equals</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>11</td>
<td>0.1225</td>
<td>-1.005</td>
<td>1 1 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.0875</td>
<td>-0.746</td>
<td>4 3 2 1 1 0 0</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>0.0625</td>
<td>-0.570</td>
<td>11 8 6 4 3 2 2</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>0.0453</td>
<td>-0.461</td>
<td>20 16 12 10 8 6 5</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0.0355</td>
<td>-0.383</td>
<td>33 28 23 20 17 14 12</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0252</td>
<td>-0.326</td>
<td>45 41 36 32 29 26 23</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0.0185</td>
<td>-0.219</td>
<td>59 55 51 47 43 40 37</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.0159</td>
<td>-0.137</td>
<td>79 67 63 60 57 54 51</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.1192</td>
<td>-1.278</td>
<td>2 1 1 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.0904</td>
<td>-1.012</td>
<td>5 3 2 1 1 0 0</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.0696</td>
<td>-0.802</td>
<td>10 7 5 4 3 2 1</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0.0519</td>
<td>-0.634</td>
<td>18 14 11 8 6 5 4</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0.0390</td>
<td>-0.498</td>
<td>25 23 19 16 13 11 9</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>0.0290</td>
<td>-0.387</td>
<td>40 35 30 26 22 19 17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.0215</td>
<td>-0.288</td>
<td>51 46 41 37 33 30 27</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.0159</td>
<td>-0.234</td>
<td>62 57 53 49 45 41 38</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.0113</td>
<td>-0.185</td>
<td>73 67 63 59 55 51 48</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.0082</td>
<td>-0.137</td>
<td>84 78 74 70 66 62 59</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0.0062</td>
<td>-0.093</td>
<td>95 88 84 80 76 72 69</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>0.0859</td>
<td>-0.764</td>
<td>20 16 12 9 7 5 4</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.0618</td>
<td>-0.608</td>
<td>29 25 21 17 13 10 8</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0.0327</td>
<td>-0.406</td>
<td>38 32 28 23 20 17 14</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.0257</td>
<td>-0.320</td>
<td>48 42 37 32 29 25 22</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.0206</td>
<td>-0.243</td>
<td>57 52 47 42 38 35 31</td>
</tr>
</tbody>
</table>

Expected no. of random MSPs with score at least \( S \): 50 9 2 0.3 0.06 0.01 0.002
Automating Blast

- **BioPerl**
  - Provides API’s for RemoteBlast and LocalBlast.
  - Provides API’s for parsing Blast output.
    - Bio::Search::HSP::GenericHSP
    - Bio::Search::Hit::BlastHit

- **Database used** – swissprot (UniprotKB)
Automating Blast

#Do a remote blast on swissprot database
Bio::Tools::Run::RemoteBlast->new('-prog' => 'blastp',
    '-data' => 'swissprot',
    '-expect' => '1e-10');

#Submit input to blast
$blastoutput = submit_blast($input);

#For each output match obtained
#Select the desired blast parameters
#If the number of outputs is greater than 5, exit
Blast output

► Input

► Output
Importing Blast Data Into MySQL Database

Vikas Sharma
### SQL Query:

```sql
SELECT * FROM `blastresults` LIMIT 0, 30
```

### Results:

<table>
<thead>
<tr>
<th>QueryId</th>
<th>EstId</th>
<th>Time</th>
<th>Genoid</th>
<th>GeneName</th>
<th>Description</th>
<th>Score</th>
<th>Gaps</th>
<th>Identities</th>
<th>E_Value</th>
<th>QueryLength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>03:52</td>
<td>Q10453</td>
<td>spG10453.2HG31_CAEEL</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>334</td>
<td>0</td>
<td>100</td>
<td>1e-30</td>
<td>66</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
<td>03:52</td>
<td>P64245</td>
<td>spP64245.2HG33_RAT</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>334</td>
<td>0</td>
<td>100</td>
<td>1e-30</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>03:52</td>
<td>Q0XW51</td>
<td>spQ0XW51.3HG33_TRIP5</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>331</td>
<td>0</td>
<td>100</td>
<td>99.5294117847069</td>
<td>3e-30</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>03:52</td>
<td>Q1U2B1</td>
<td>spQ1U2B1.3HG33_CAEEL</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>331</td>
<td>0</td>
<td>100</td>
<td>99.5294117847069</td>
<td>3e-30</td>
</tr>
<tr>
<td>5</td>
<td>78</td>
<td>03:52</td>
<td>Q5RCC5</td>
<td>spQ5RCC5.3HG33_PONAB</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>330</td>
<td>0</td>
<td>100</td>
<td>99.5294117847069</td>
<td>4e-30</td>
</tr>
<tr>
<td>6</td>
<td>78</td>
<td>03:52</td>
<td>Q5RCC5</td>
<td>spQ5RCC5.3HG33_PONAB</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>330</td>
<td>0</td>
<td>100</td>
<td>99.5294117847069</td>
<td>4e-30</td>
</tr>
<tr>
<td>7</td>
<td>78</td>
<td>03:52</td>
<td>Q0XW51</td>
<td>spQ0XW51.3HG33_TRIP5</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>219</td>
<td>0</td>
<td>100</td>
<td>99.6512739130436</td>
<td>8e-16</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
<td>03:52</td>
<td>Q64245</td>
<td>spQ64245.2HG33_RAT</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>219</td>
<td>0</td>
<td>100</td>
<td>99.6512739130436</td>
<td>8e-16</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>03:52</td>
<td>Q64245</td>
<td>spQ64245.3HG2_GRILDO</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>216</td>
<td>0</td>
<td>100</td>
<td>93.4792666395652</td>
<td>2e-17</td>
</tr>
<tr>
<td>10</td>
<td>78</td>
<td>03:52</td>
<td>Q402E2</td>
<td>spQ402E2.3HG33_AILLO</td>
<td>ReName: Full=H3 Histone H3 type 1</td>
<td>216</td>
<td>0</td>
<td>100</td>
<td>93.4792666395652</td>
<td>2e-17</td>
</tr>
<tr>
<td>11</td>
<td>60</td>
<td>03:52</td>
<td>No Match</td>
<td>No Match</td>
<td>No Match</td>
<td>0</td>
<td>0</td>
<td>No Match</td>
<td>No Match</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
<td>03:52</td>
<td>Q32L27</td>
<td>spQ32L27.1UB2G2_BOVIN</td>
<td>ReName: Full=Ubiquitin-conjugating enzyme E2 Q2 type 1</td>
<td>539</td>
<td>0</td>
<td>99.1904761904762</td>
<td>3e-54</td>
<td>105</td>
</tr>
<tr>
<td>13</td>
<td>81</td>
<td>03:52</td>
<td>Q32L27</td>
<td>spQ32L27.1UB2G2_BOVIN</td>
<td>ReName: Full=Ubiquitin-conjugating enzyme E2 Q2 type 1</td>
<td>539</td>
<td>0</td>
<td>99.1904761904762</td>
<td>3e-54</td>
<td>105</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>03:52</td>
<td>Q7YQJ9</td>
<td>spQ7YQJ9.1UB2G2_RABBIT</td>
<td>ReName: Full=Ubiquitin-conjugating enzyme E2 Q2 type 1</td>
<td>539</td>
<td>0</td>
<td>99.1904761904762</td>
<td>3e-54</td>
<td>105</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>03:52</td>
<td>Q9WAN8</td>
<td>spQ9WAN8.1UB2G2_HUMAN</td>
<td>ReName: Full=Ubiquitin-conjugating enzyme E2 Q2 type 1</td>
<td>539</td>
<td>0</td>
<td>99.1904761904762</td>
<td>3e-54</td>
<td>105</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>03:52</td>
<td>Q10453</td>
<td>spQ10453.1UB2G2_BOVIN</td>
<td>ReName: Full=Ubiquitin-conjugating enzyme E2 Q2 type 1</td>
<td>539</td>
<td>0</td>
<td>99.1904761904762</td>
<td>3e-54</td>
<td>105</td>
</tr>
</tbody>
</table>
Deliverables – Issues - Improvements

Nathan Helmick
Issues Currently Being Tracked

► Runtime can be lengthy - ~6 seconds per EST
  ▪ Most of the time spent waiting on results from Blast Search

► During repeated testing, it has been observed that the connection to Blast will be refused
  ▪ Appears to be related to over-using the resource
  ▪ May need to add some kind of timeout after so many searches
Deliverable Items

► MySQL database recovery script file containing all input data

► Final report containing system requirements, installation instructions, database design, and design methodology

► Perl and PHP script source code
Further Enhancements

► Establish tighter cohesion between the perl and php scripts

► Improve database storage types for easier searching

► Modify parsed data based on customer feedback / requests
References


