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.

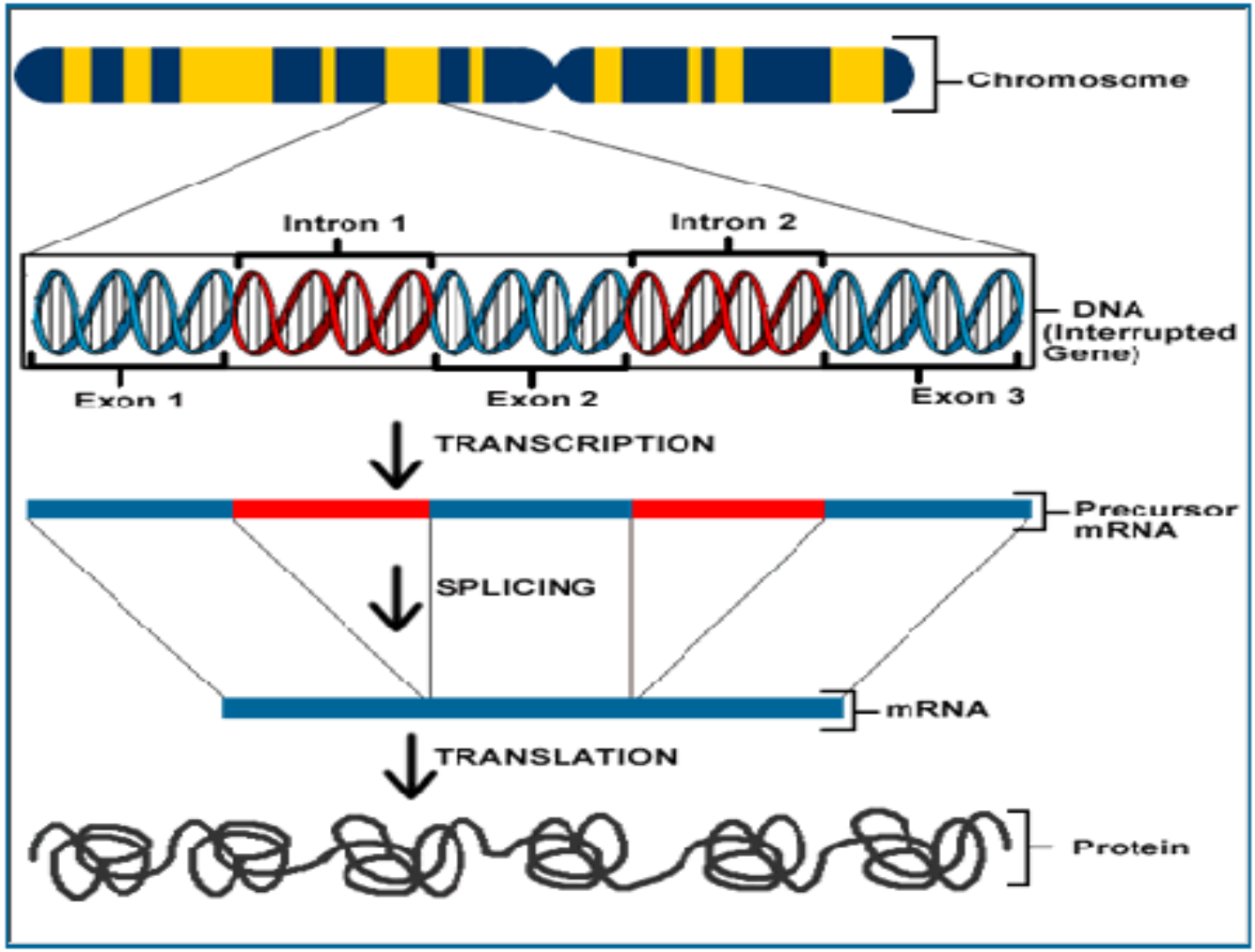
How Are They Made?

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

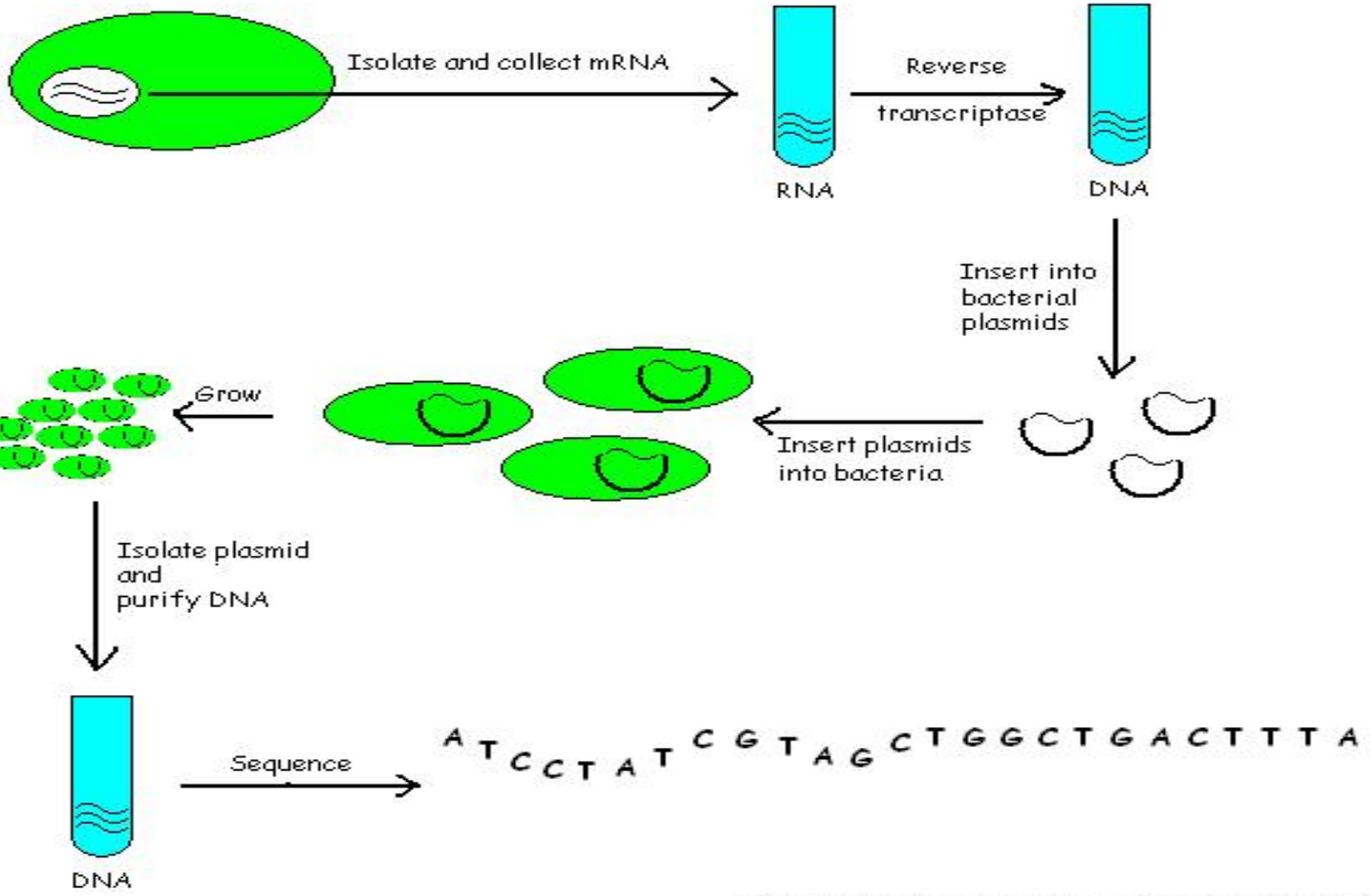
- Generated by sequencing either one or both ends of an expressed gene.

- 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

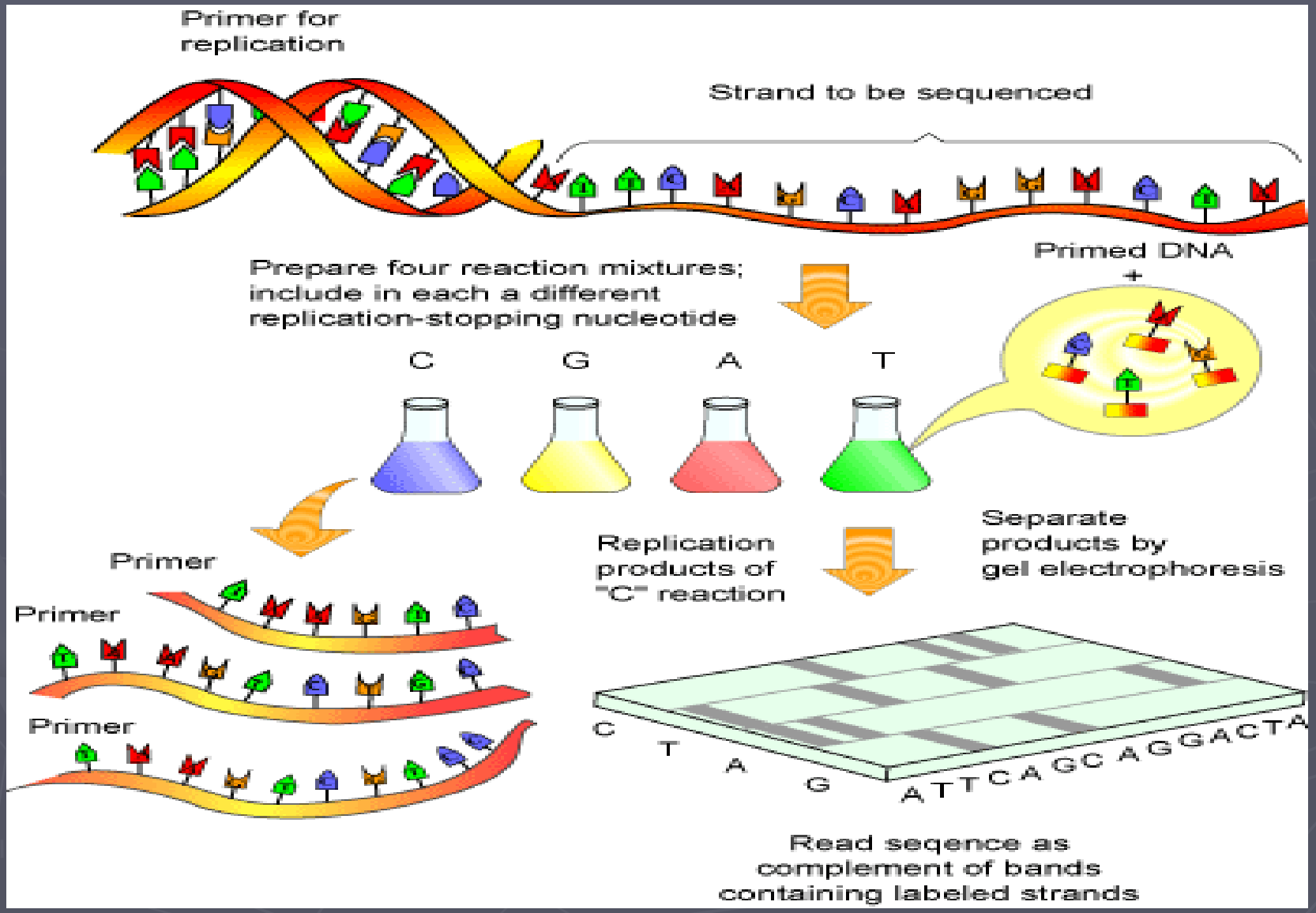


Formation of a cDNA Library

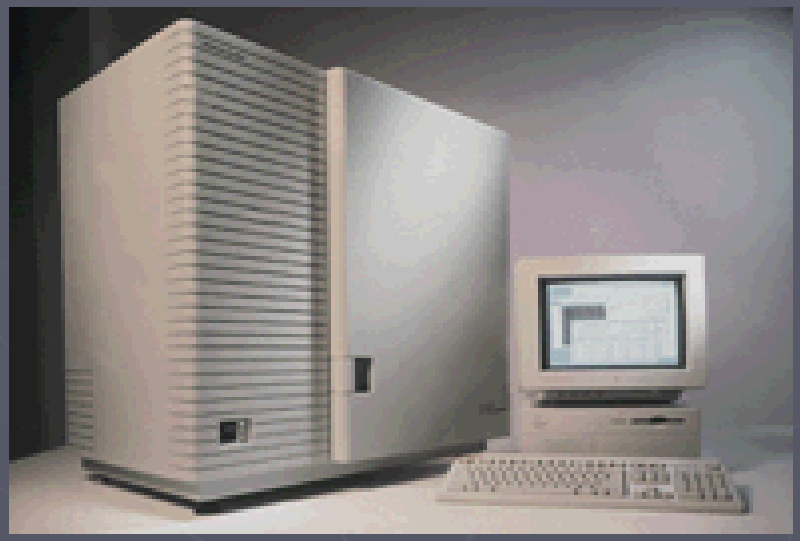


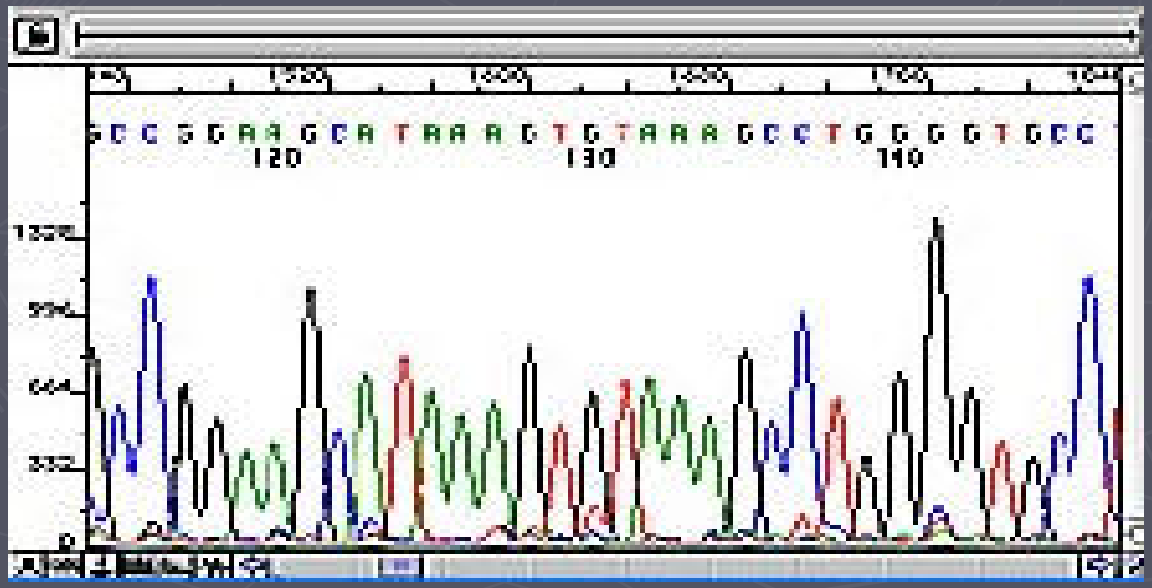
Info provided by: cLass1C and uLtRa_GoLiAtH

Chemistry behind sequencing



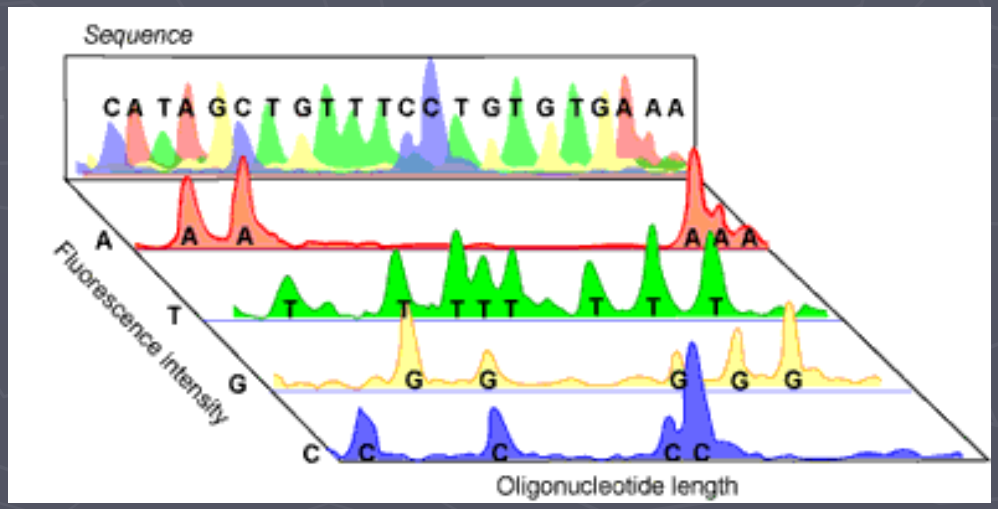
ABI 377 DNA Sequencers





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 .

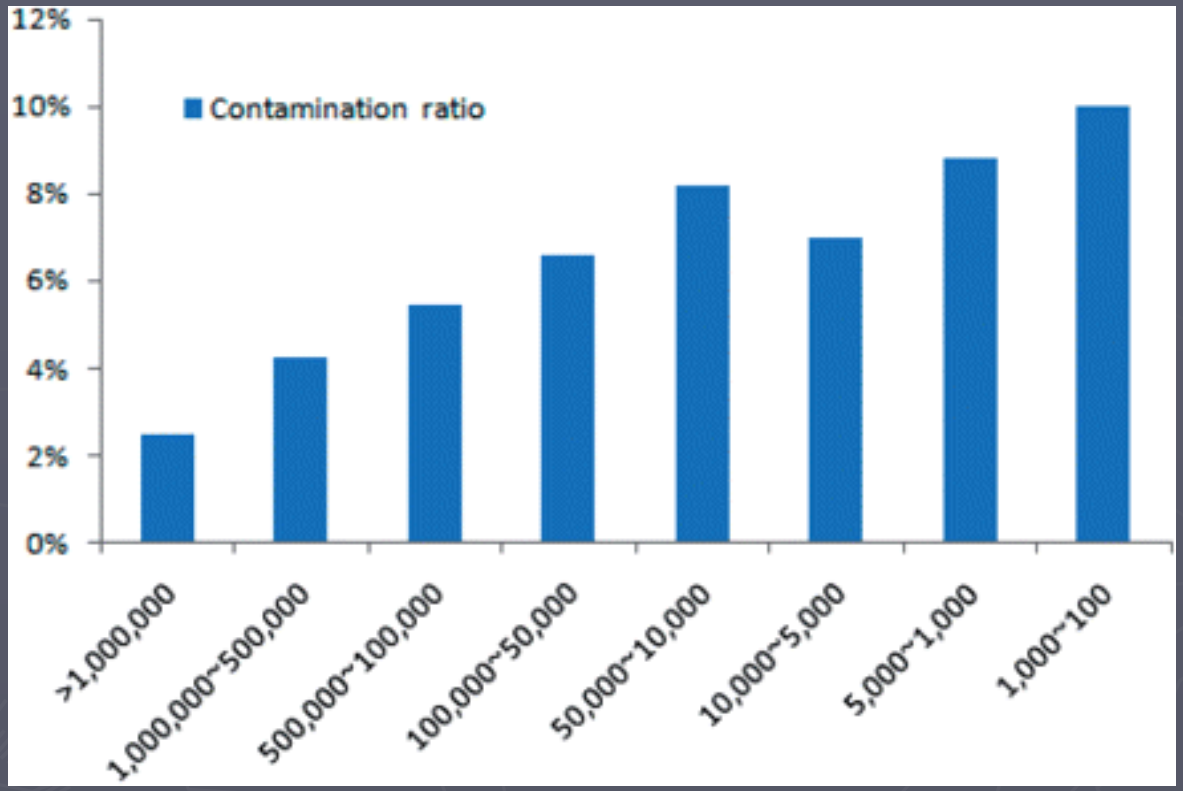


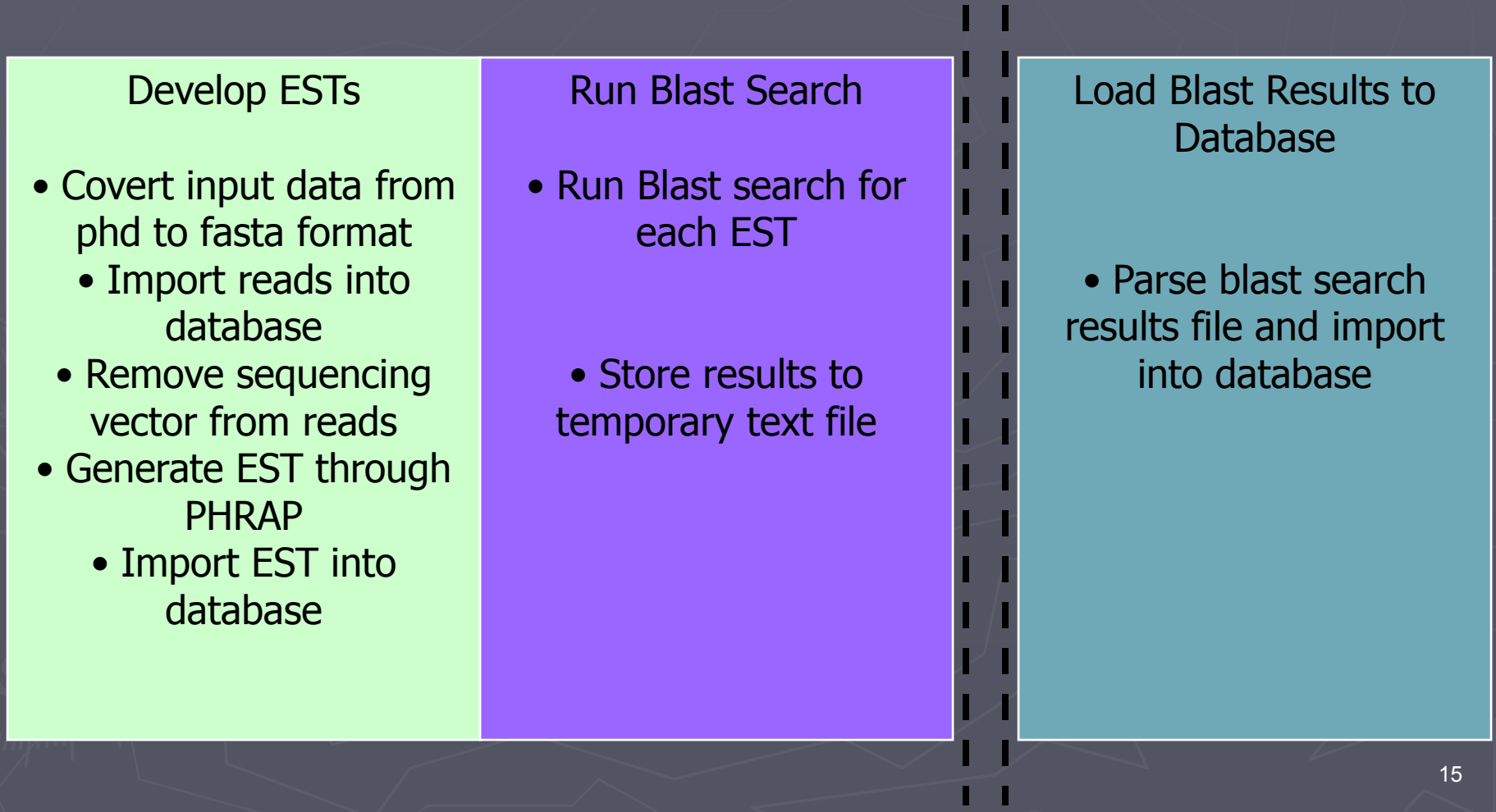
Figure 1. Distribution of average contamination ratio of sequencing centers. Sequencing centers were classified according to the number of their total sequences in dbEST and were calculated an average contamination ratio of each class. The *x*-axis represents the classes of sequencing centers and *y*-axis 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

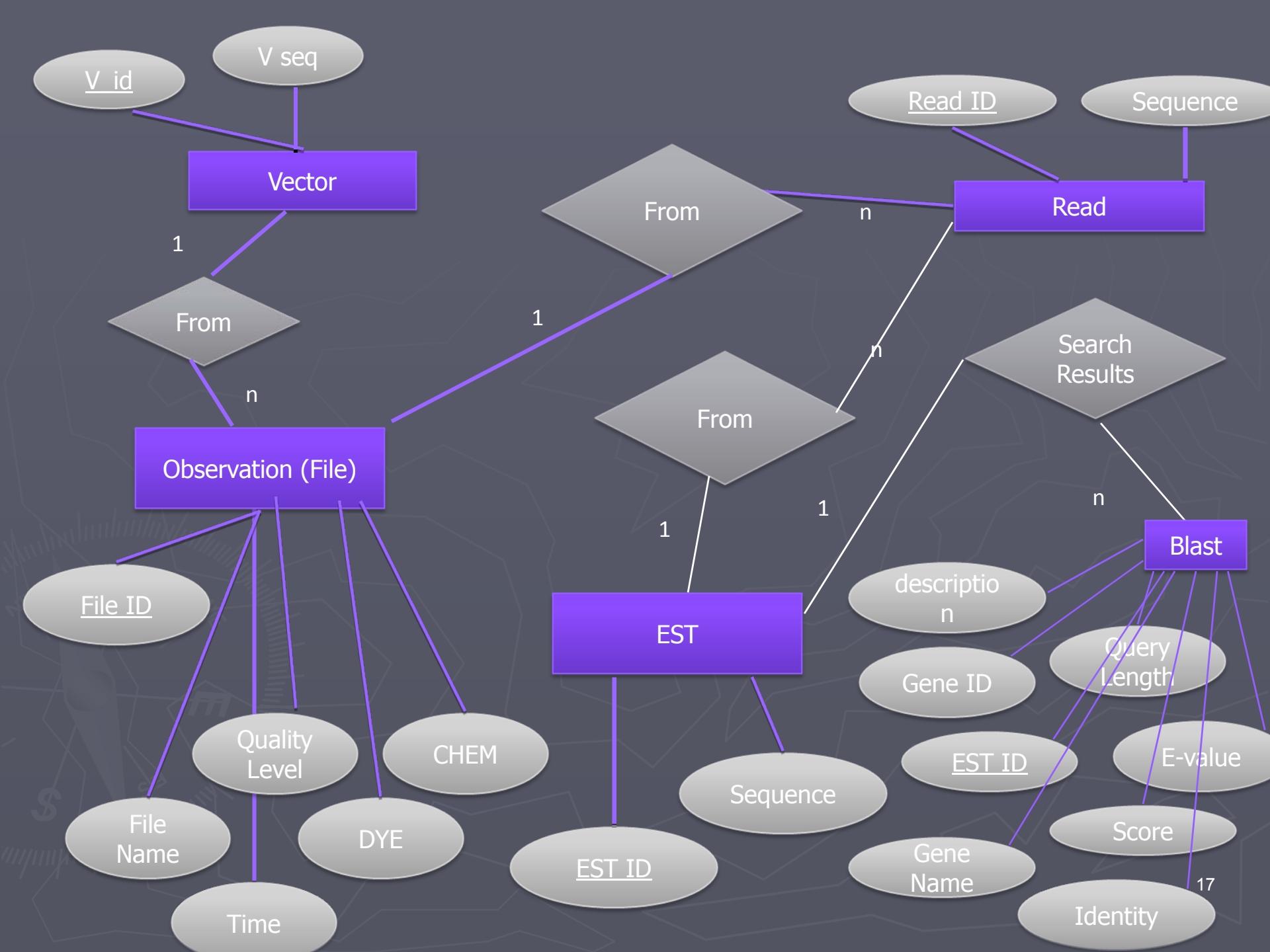
Nathan Helmick

System Flow Overview



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


V_id
V seq
Vector
1
From
n
Observation (File)
File ID
File Name
Quality Level
CHEM
DYE
Time
From
1
n
Read ID
Sequence
Read
From
1
n
EST
EST ID
Sequence
Search Results
1
n
Blast
descriptio n
Gene ID
Query Length
EST ID
E-value
Score
Gene Name
Identity

Converting Inputs To ESTs

Nathan Helmick

Coversion 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

PHD2FASTA User's Guide, 1999, Phil Green

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

Cloning Vectors, Molecular Biology Web Book – Chapter 9, 2009, http://www.web-books.com/MoBio/

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

PHRAP User's Guide, 1999, Phil Green

Cross Match

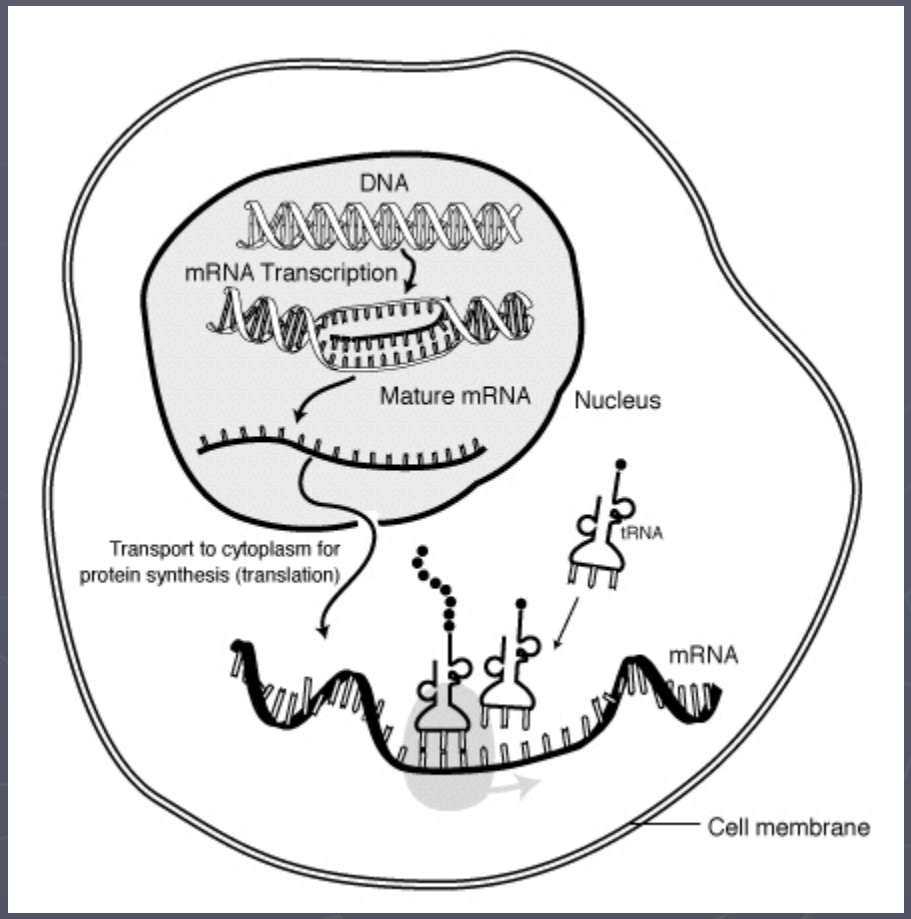
- Cross Match is the tool that we have selected for removing cloning vectors
- It accepts a vector file containing the sequence used for cloning
  - Many labs only create a single sequence file that contains all potential vectors they may use for cloning
- If segments of a sequence are determined to be from the cloning vector a series ‘x’s are placed in the effected region

PHRAP User’s Guide, 1999, Phil Green

Sequences We Recieve

- 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

mRNA



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 availalbe
- 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 math count

Dwyer, Rex. Genomic Perl. New York.
Cambridge University Press. 2002. 978-0-511-06339-8

Why Use PHRAP?

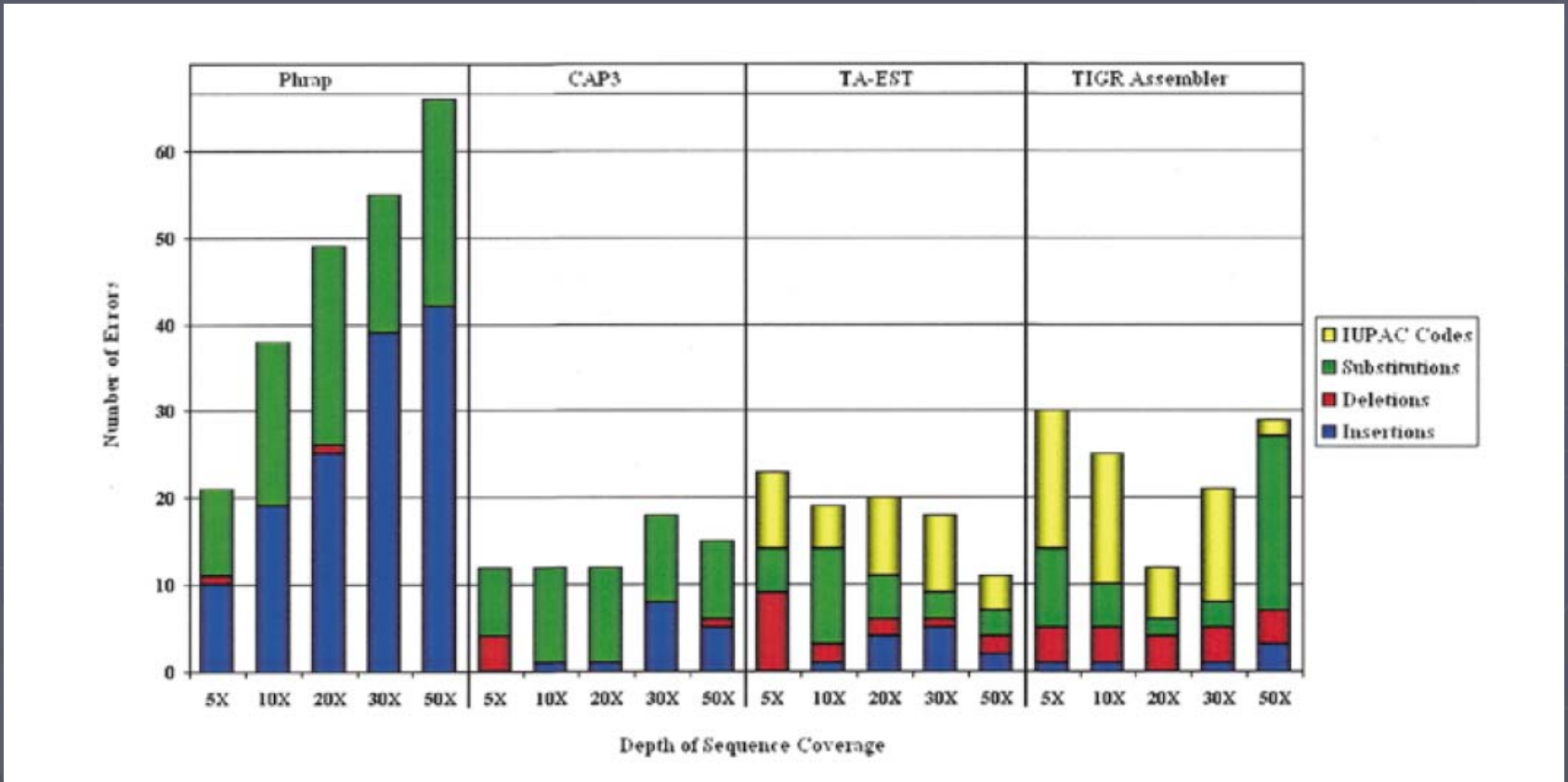
- Even though we aren't exactly reading shot-gun blasted DNA, it is possible our genes could extend the capacity of our analysis equipment
  - Current data set does not appear to be the case
- PHRAP will identify and combine repeated reads of the same gene or segments of the same gene
- Currently we 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 qualititative 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 then 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%

Liang, Feng, Holt, Ingenborg, et. al. *An optimized protocol for anlayis of EST sequences*, Nucleic Acids Research, 2000, Vol.28, No.18

Analysis of PHRAP vs Others



Liang, Feng, Holt, Ingenborg, et. al. *An optimized protocol for anlayis of EST sequences*, Nucleic Acids Research, 2000, Vol.28, No.18

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

Sample Perl

```
#!/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 = "";
}
```

Sample Perl

```
#!/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 = "";
}
```

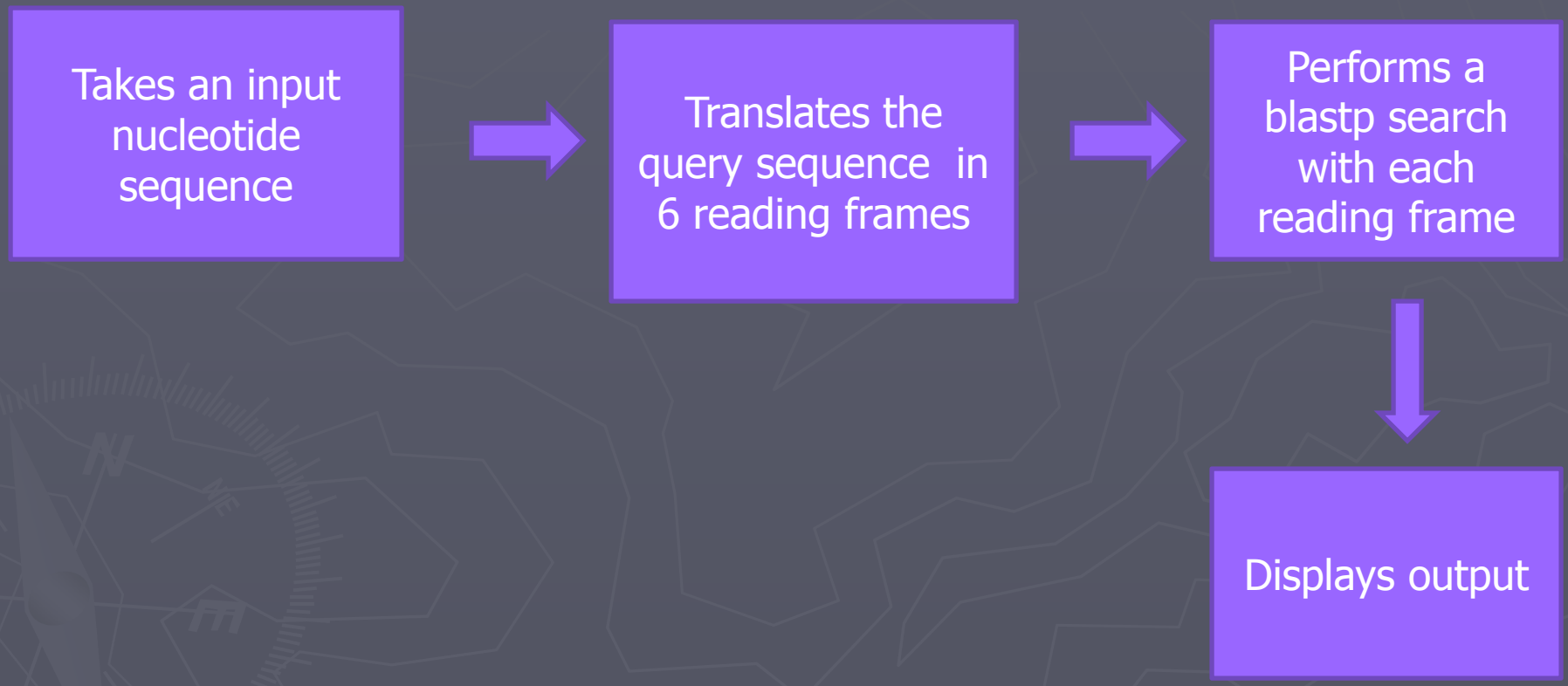
Performing BLAST Search On ESTs

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



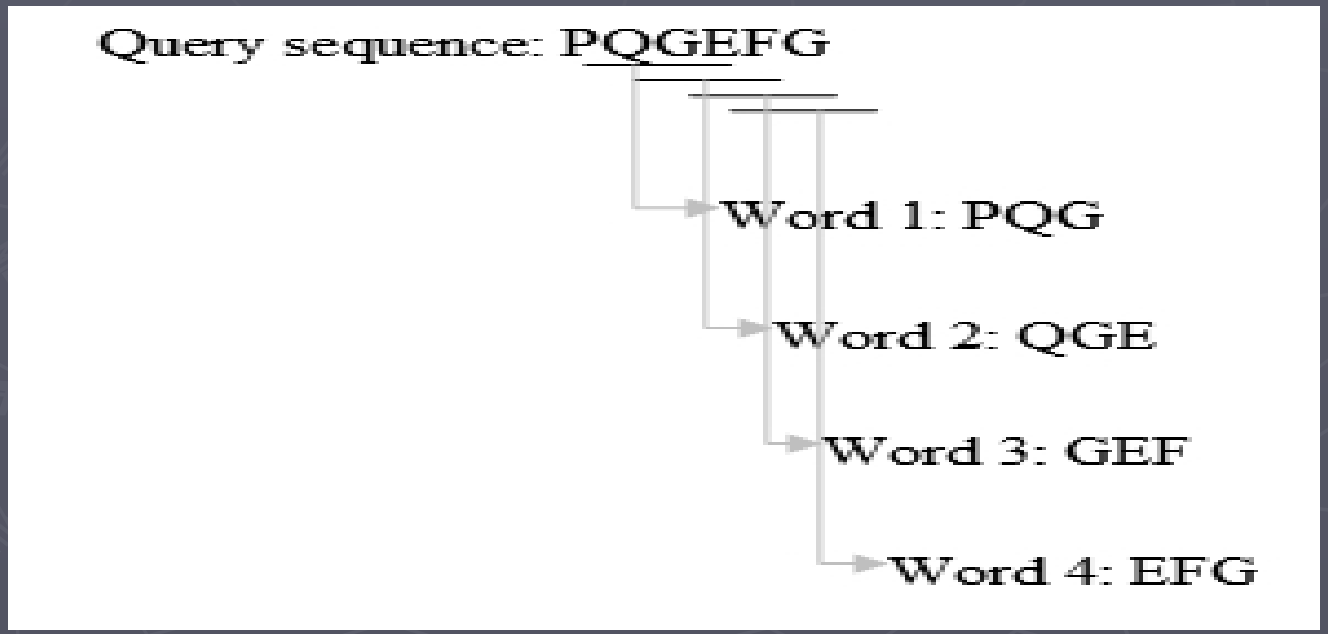
- http://blast.ncbi.nlm.nih.gov/Blast.cgi

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.
  - PPCDPPPPPKDKKKKDDGPP
  - AAATAAAAAAAATAAAAAAT
- 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.

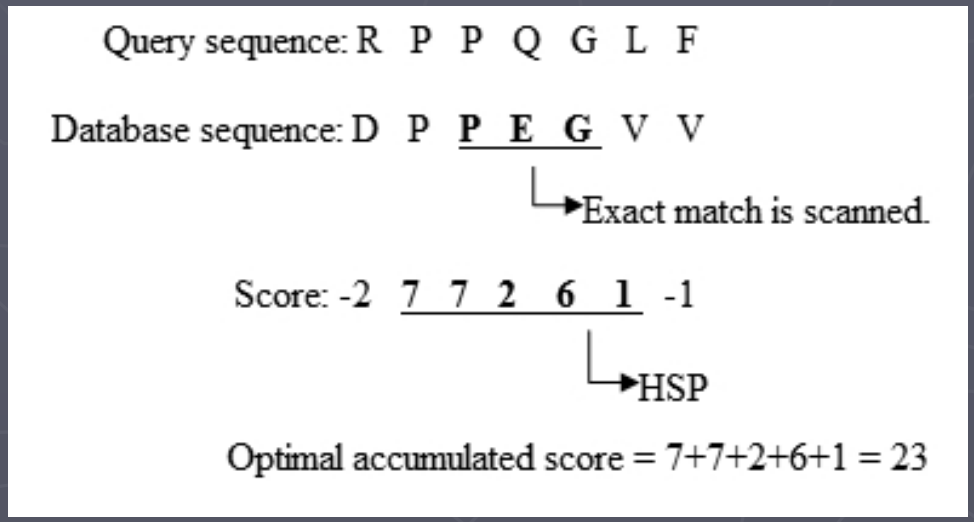


How blastp works?

- 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?

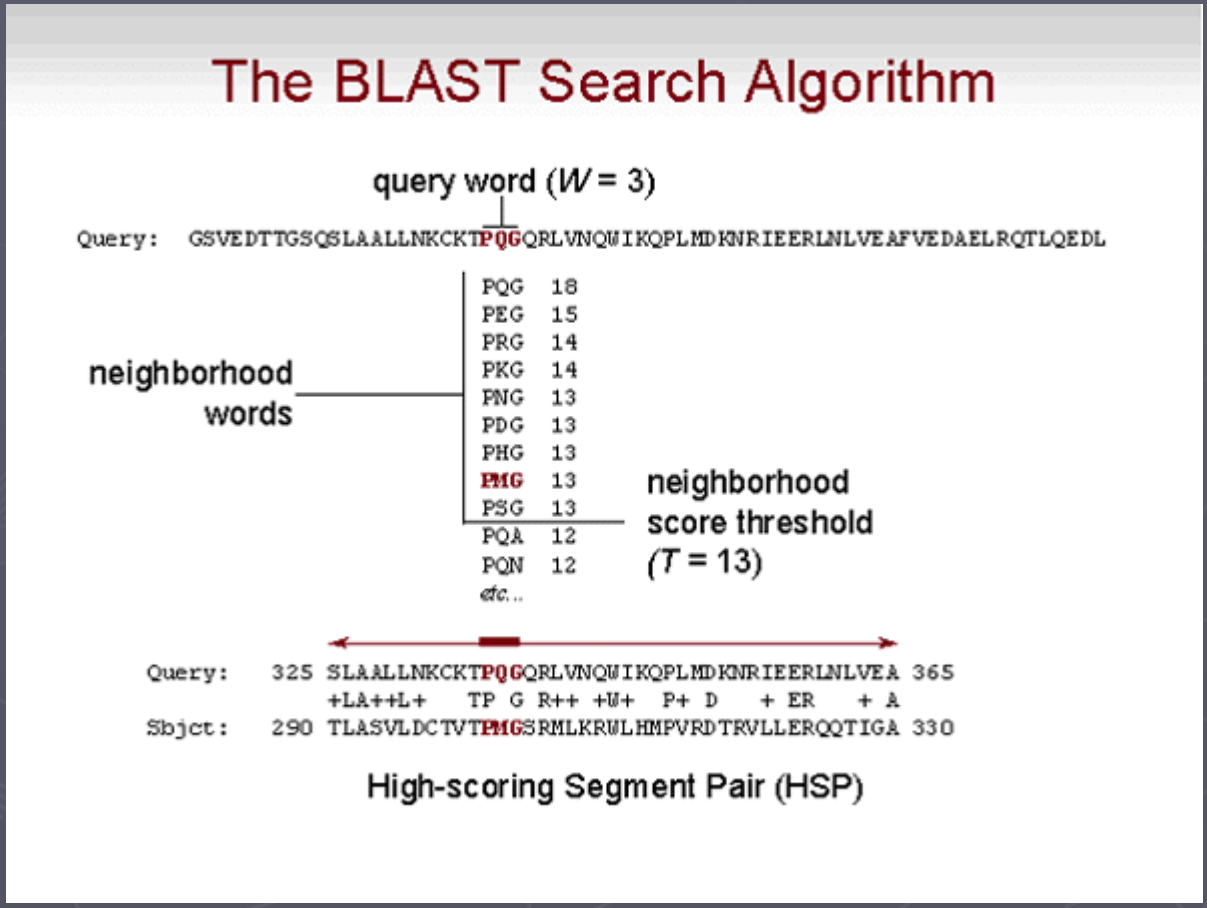
- Extend the exact matches to high-scoring segment pair (HSP).



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?



BLAST Statistics

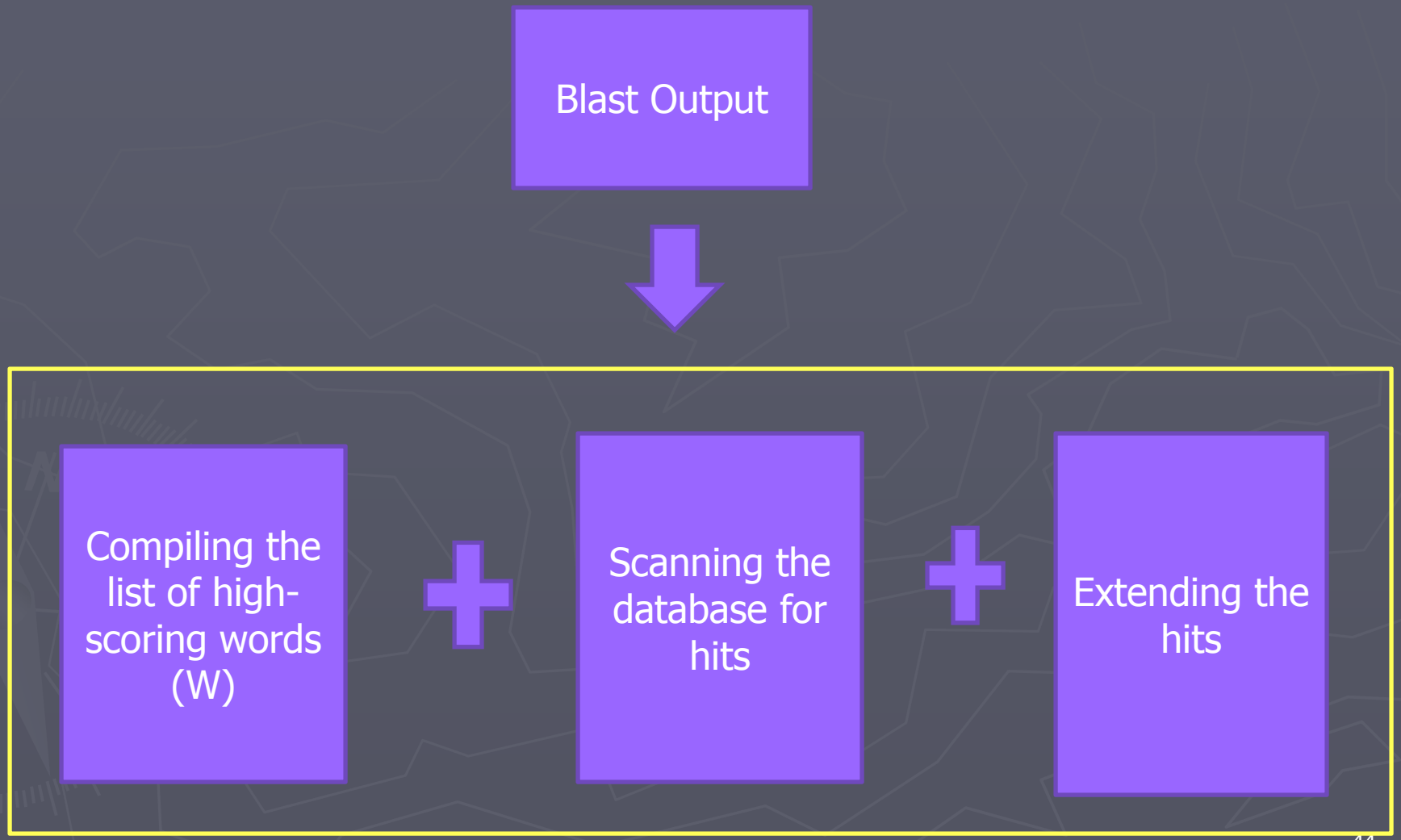
- Approx. 50 words are found for every residue in a protein sequence.
  - So , for a sequence of length 250 the total number of words will be 250 * 50 = 12500
- Probability of finding a sequence having score>=S

$1 - e^{-y}$, where $y = Kmn\ e^{-\lambda S}$

m , n = length of the input sequences.

K , λ = parameters of statistical significance.

BLAST Statistics



WordList vs CPU time

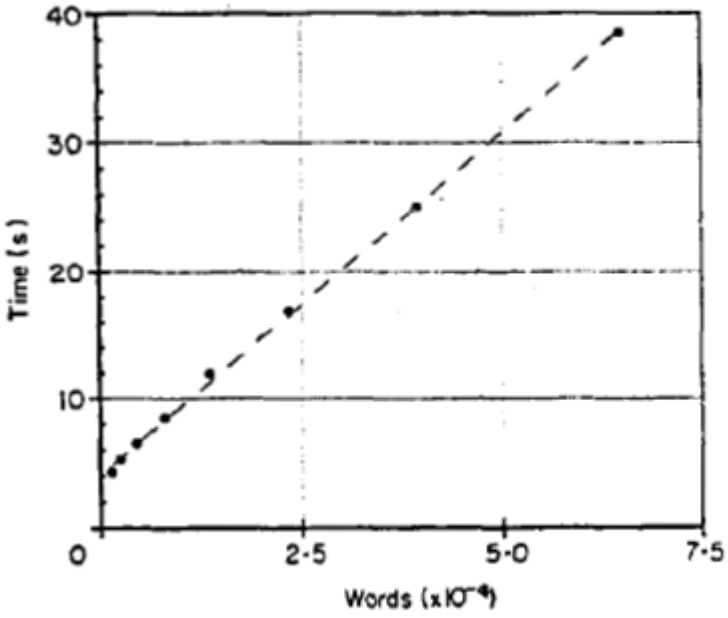


Figure 2. The central processing unit time required to execute BLAST on the PIR protein database (Release 23·0) as a function of the size of the word list generated.

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

w	T	Probability of a hit $\times 10^5$	Linear regression $-\ln(q) = aS + b$		Implied % of MSPs missed by BLAST when S equals						
			a	b	45	50	55	60	65	70	75
3	11	253	0·1236	−1·005	1	1	0	0	0	0	0
	12	147	0·0875	−0·746	4	3	2	1	1	0	0
	13	83	0·0625	−0·570	11	8	6	4	3	2	2
	14	48	0·0463	−0·461	20	16	12	10	8	6	5
	15	26	0·0328	−0·353	33	28	23	20	17	14	12
	16	14	0·0232	−0·263	46	41	36	32	29	26	23
	17	7	0·0158	−0·191	59	55	51	47	43	40	37
	18	4	0·0109	−0·137	70	67	63	60	57	54	51
4	13	127	0·1192	−1·278	2	1	1	0	0	0	0
	14	78	0·0904	−1·012	5	3	2	1	1	0	0
	15	47	0·0686	−0·802	10	7	5	4	3	2	1
	16	28	0·0519	−0·634	18	14	11	8	6	5	4
	17	16	0·0390	−0·498	28	23	19	16	13	11	9
	18	9	0·0290	−0·387	40	35	30	26	22	19	17
	19	5	0·0215	−0·298	51	46	41	37	33	30	27
	20	3	0·0159	−0·234	62	57	53	49	45	41	38
5	15	64	0·1137	−1·525	3	2	1	1	0	0	0
	16	40	0·0882	−1·207	6	4	3	2	1	1	0
	17	25	0·0679	−0·939	12	9	6	4	3	2	2
	18	15	0·0529	−0·754	20	15	12	9	7	5	4
	19	9	0·0413	−0·608	29	23	19	15	13	10	8
	20	5	0·0327	−0·506	38	32	28	23	20	17	14
	21	3	0·0257	−0·420	48	42	37	32	29	25	22
	22	2	0·0200	−0·343	57	52	47	42	38	35	31
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

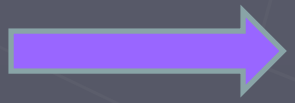
```
#Do a remote blast on swissprot database
Bio::Tools::Run::RemoteBlast->new('-prog' => 'blastp',
                                  '-data' => 'swissprot',
                                  '-expect' => '1e-10');

#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

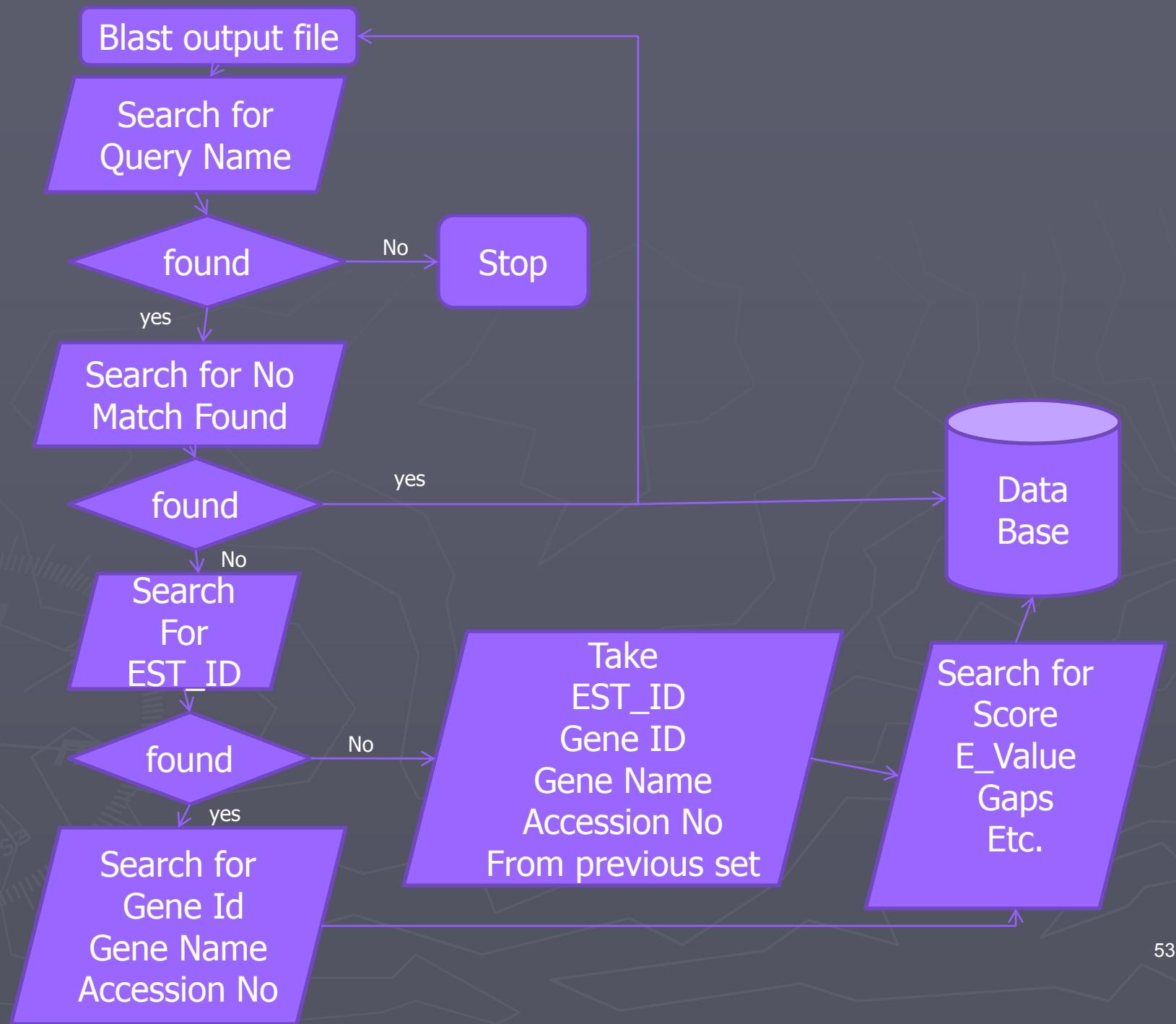
Importing Blast Data Into MySQL Database

Vikas Sharma



BLAST

```
Query Name: fasta.screen.Contig8
        No Matches Found ...

Query Name: fasta.screen.Contig9
               EST_ID = 159
               name = sp|P56384.1|AT5G3_MOUSE
               description = RecName: Full=ATP synthase lipid-binding protein, mitochondrial; AltName: Full=ATP synthase proteolipid P3; AltNar
               accession = P56384
               score = 239
               Algorithm = BLASTX
               Expect Value = 3e-32
               Fraction Identical = 0.646017699115044
               Fraction conserved = 0.769911504424779
               Gaps = 1
               query sequence = MYSCAKFVSCPAVVRSTSRTFLRPMSASVFSRPEIQ-NEQAQLLPAPRNALVQTVRRDLQTSIASRDIDTAAKFigagaatvgvagsgagigtvFGSLIIGYARNPSLKTA
               Percent Identity = 64.6017699115044
               Hit String = MFACAKLARTPALIRAGSRVAYRPISASVLSRPETRTGEGSTVFNGAQNGVCQLIRREFQTSVISRDIDTAAKFIGAGAATVGVAGSGAGIGTVFGSLIIGYARNPSLKQQLF
               Homology Sequence = M++CAK    PA++R+ SR   RP+SASV SRPE +  E + +    +N + Q +RR+ QTS+ SRDIDTAAKFIGAGAATVGVAGSGAGIGTVFGSLIIGYARNPSL
               Query length = 113
               Sequence lenth = 113
               rank = 1

               score = 155
               Algorithm = BLASTX
               Expect Value = 3e-32
               Fraction Identical = 0.96969696969697
               Fraction conserved = 0.96969696969697
               Gaps = 0
               query sequence = KQQLFSYAILGFALSEAMGLFCLTVAFLILFAM
               Percent Identity = 96.969696969697
               Hit String = KQQLFSYAILGFALSEAMGLFCLMVAFLILFAM
               Homology Sequence = KQQLFSYAILGFALSEAMGLFCL VAFLILFAM
               Query length = 33
               Sequence lenth = 33
               rank = 2

               EST_ID = 159
               name = sp|Q5RFL2.1|AT5G3_PONAB
               description = RecName: Full=ATP synthase lipid-binding protein, mitochondrial; AltName: Full=ATP synthase proteolipid P3; AltNar
               accession = Q5RFL2
               score = 225
               Algorithm = BLASTX
               Expect Value = 1e-30
               Fraction Identical = 0.62280701754386
               Fraction conserved = 0.745614035087719
               Gaps = 2
               query sequence = MYSCAKFVSCPAVVRSTSRTFLRPMSASVFSRPEIQ--NEQAQLLPAPRNALVQTVRRDLQTSIASRDIDTAAKFigagaatvgvagsgagigtvFGSLIIGYARNPSLKT
               Percent Identity = 62.280701754386
               Hit String = MFACAKLACTPSLIRAGSRVAYRPISASVLSRPEASRTGEGSAVFNGAQNGVSQLIQREFQTSAISRDIDTAAKFIGAGAATVGVAGSGAGIGTVFGSLIIGYARNPSLKQQLF
               Homology Sequence = M++CAK    P+++R+ SR   RP+SASV SRPE     E + +    +N + Q ++R+ QTS  SRDIDTAAKFIGAGAATVGVAGSGAGIGTVFGSLIIGYARNPS
               Query length = 114
               Sequence lenth = 114
               rank = 1

               score = 155
               Algorithm = BLASTX
               Expect Value = 1e-30
               Fraction Identical = 0.96969696969697
               Fraction conserved = 0.96969696969697
               Gaps = 0
               query sequence = KQQLFSYAILGFALSEAMGLFCLTVAFLILFAM
               Percent Identity = 96.969696969697
               Hit String = KQQLFSYAILGFALSEAMGLFCLMVAFLILFAM
               Homology Sequence = KQQLFSYAILGFALSEAMGLFCL VAFLILFAM
               Query length = 33
               Sequence lenth = 33
               rank = 2

               EST_ID = 159
```


Blast output file
Search for Query Name
found
No
Stop
yes
Search for No Match Found
found
yes
Data Base
No
Search For EST_ID
found
No
Take EST_ID Gene ID Gene Name Accession No From previous set
Search for Score E_Value Gaps Etc.
yes
Search for Gene Id Gene Name Accession No

Server: localhost ▸ Database: est ▸ Table: blastresults

Browse | Structure | SQL | Search | Insert | Export | Import | Operations | Empty | Drop

Database

est (5)

est (5)

- blastresults
- est
- inputread
- readtoest
- vectors

Showing rows 0 - 29 (65 total, Query took 0.0008 sec)

SQL query:

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

Profiling [Edit] [Explain SQL] [Create PHP Code] [Refresh]

Show: 30 row(s) starting from record # 30 > >> Page number: 1

in horizontal mode and repeat headers after 100 cells

Sort by key: None

←T→	QueryId	EstId	Time	GeneId	GeneName	Description	Score	Gaps	Identities	E_Value	QueryLength
	1	78	03:52:50	Q10453	sp|Q10453.2|H331_CAEEL	RecName: Full=Histone H3.3 type 1	334	0	100	1e-30	68
	2	78	03:52:50	P84245	sp|P84245.2|H33_RAT	RecName: Full=Histone H3.3sp|P84243.2|H33_HUMAN R...	334	0	100	1e-30	68
	3	78	03:52:50	Q8WSF1	sp|Q8WSF1.3|H33_TRIPS	RecName: Full=Histone H3.3	331	0	98.5294117647059	3e-30	68
	4	78	03:52:50	Q9U281	sp|Q9U281.3|H332_CAEEL	RecName: Full=Histone H3.3 type 2	331	0	98.5294117647059	3e-30	68
	5	78	03:52:50	Q5RCC9	sp|Q5RCC9.3|H33_PONAB	RecName: Full=Histone H3.3	330	0	98.5294117647059	4e-30	68
	6	79	03:52:50	Q5RCC9	sp|Q5RCC9.3|H33_PONAB	RecName: Full=Histone H3.3	219	0	95.6521739130435	8e-18	46
	7	79	03:52:50	Q8WSF1	sp|Q8WSF1.3|H33_TRIPS	RecName: Full=Histone H3.3	219	0	95.6521739130435	8e-18	46
	8	79	03:52:50	P84245	sp|P84245.2|H33_RAT	RecName: Full=Histone H3.3sp|P84243.2|H33_HUMAN R...	219	0	95.6521739130435	8e-18	46
	9	79	03:52:50	Q64400	sp|Q64400.3|H32_CRILO	RecName: Full=Histone H3.2	216	0	93.4782608695652	2e-17	46
	10	79	03:52:50	Q402E2	sp|Q402E2.3|H33A_LILLO	RecName: Full=Histone H3.3a; AltName: Full=Somati...	216	0	93.4782608695652	2e-17	46
	11	80	03:52:50	No Match	No Match	No Match	0	0	No Match	No Match	0
	12	81	03:52:50	Q32L27	sp|Q32L27.1|UB2Q2_BOVIN	RecName: Full=Ubiquitin-conjugating enzyme E2 Q2;...	539	0	96.1904761904762	3e-54	105
	13	81	03:52:50	Q8K2Z8	sp|Q8K2Z8.2|UB2Q2_MOUSE	RecName: Full=Ubiquitin-conjugating enzyme E2 Q2;...	539	0	96.1904761904762	3e-54	105
	14	81	03:52:50	Q7YQJ9	sp|Q7YQJ9.1|UB2Q2_RABIT	RecName: Full=Ubiquitin-conjugating enzyme E2 Q2;...	539	0	96.1904761904762	3e-54	105
	15	81	03:52:50	Q8WVN8	sp|Q8WVN8.1|UB2Q2_HUMAN	RecName: Full=Ubiquitin-conjugating enzyme E2 Q2;...	539	0	96.1904761904762	3e-54	105
	16	81	03:52:50	Q7TSS2	sp|Q7TSS2.2|UB2Q1_MOUSE	RecName: Full=Ubiquitin-conjugating	531	0	94.2857142857143	2e-53	105

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

Refrences

1. Byungwook Lee and Gwangsik Shin. CleanEST: a database of cleansed EST libraries. Nucleic Acids Research, 2009, Vol. 37

2. Andreas D. Baxevanis. BIOINFORMATICS : A Practical Guide to the Analysis of Genes and Proteins. WILEY PUBLICATION.

3. Helmut Kae. GENOME PROJECTS: UNCOVERING THE BLUEPRINTS OF BIOLOGY. August 2003.

4. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers , David J. Lipman : Basic Local Alignment Search Tool