Introduction to Bioinformatics
3. **DNA editing and contig assembly**

Benjamin F. Matthews
United States Department of Agriculture
Soybean Genomics and Improvement Laboratory
Beltsville, MD 20708
matthewb@ba.ars.usda.gov

What we will cover today

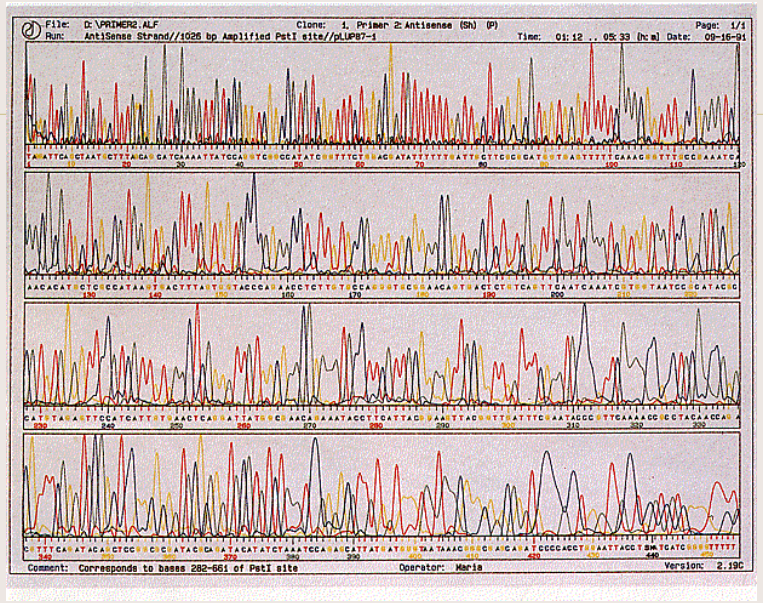
- DNA editing
  - Phred
- Sequence assembly (Contig building)
  - Phrap
  - Consed
  - CAP3
  - DNA Star - commercial software
  - http://www.phrap.org/

What we will cover today

- DNA Sequencing software
- DNA sequence assembly
- Similarity searching with a DNA sequence
- BLAST

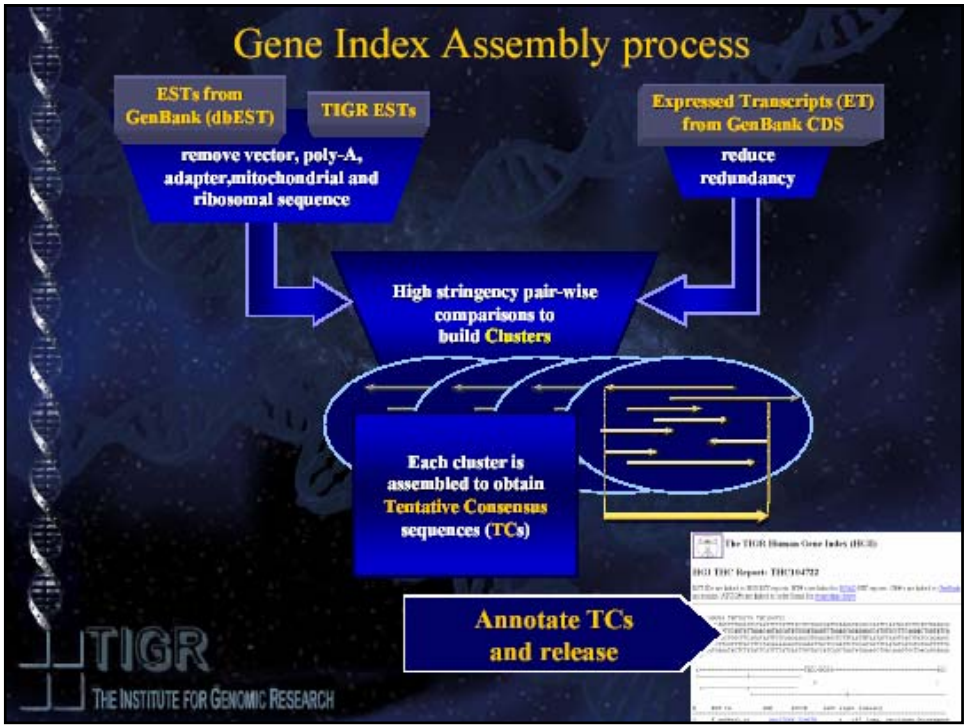
You cloned a cDNA

- Isolated mRNA
- Reverse transcribed
- Placed into vector
- Transformed and grew bacteria
- Harvested plasmid
- Sequenced insert



DNA sequence analysis

- Is full-length cDNA cloned?
- What are its properties
- What is function of encoded protein?
- Are there family members?
- Is it cloned from other organisms?



DNA sequence analysis

- DNA sequencing software
  - Phred
    - Reads DNA sequencer trace files, calls bases, assigns quality values to each called base
    - http://www.genome.washington.edu/UWGC/analysistools/phred.cfm (for Phred, Phrap, Consed)
  - Phrap
    - A program for assembling shotgun DNA sequence data into contig sequence; provides consensus quality estimates
  - Consed/Autofinish
    - A tool for viewing, editing, and finishing sequence assemblies
  - CAP3 Assembly program
    - Sequence assembly
    - http://genome.cs.mtu.edu/

DNA editing and sequence assembly - building contiguous sequences (contigs)

- Phil Green
- Genome Sciences, University of Washington

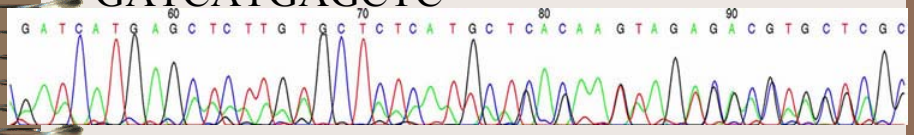
- http://www.phrap.org
- Provides software and documentation

Phred

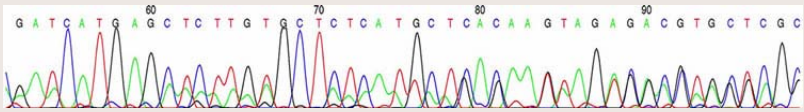
- Software reads sequencing trace files
- Calls bases
- Assigns a quality value to each called base
  - Correct and incorrect base calls
  - Quality values allow sequence trimming
- Works with Amersham Biosciences, Applied Biosystems, Beckman, LI-COR Life Sciences instruments

Which base reads are reliable

- GATCATGAGCTC



Phred



- Vector sequences must be trimmed from both ends
- Poor quality bases must be edited
- PolyA tail indicates 3’ end
- PolyT tail indicates 3’ end reverse sequence

5’ end
TTTATC**ATG**GCTGCCCCTAGGGGCGAT
GAA**TGA**TCGTATGCCAGCTAAAAAAA
AAAATCCGCCG
3’ end

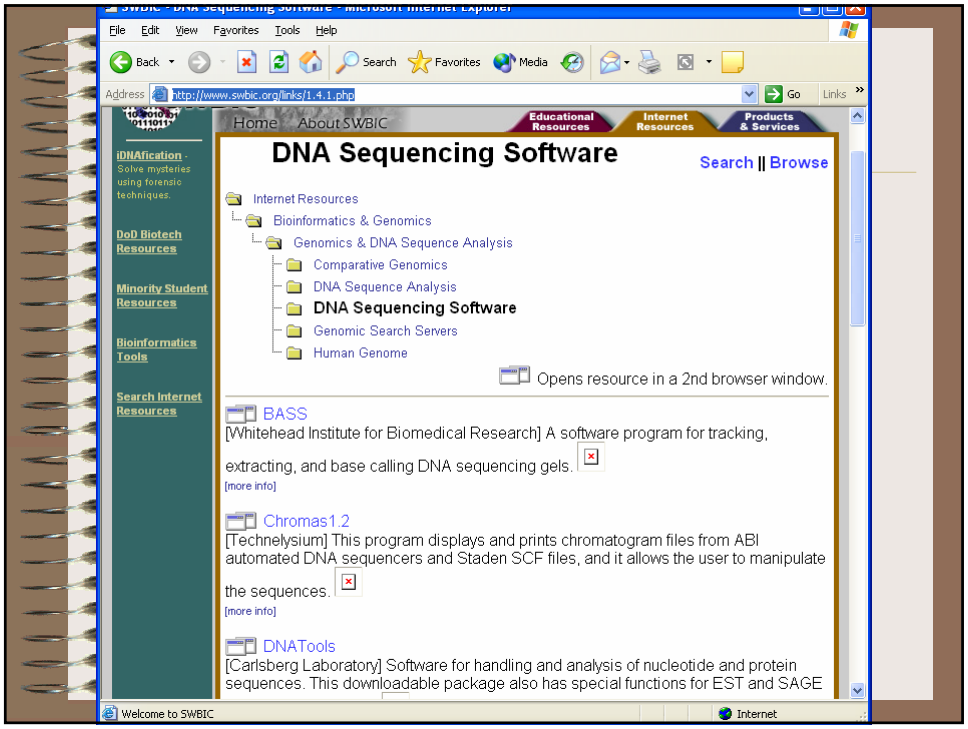
From 5’ end:
ATG = methionine - possible start site
TGA=STOP site
AAAAA…= possible polyA tail
Remaining 3’ sequence may be cloning vector sequence

Phrap

- Assembling shotgun DNA sequence data
- Improves assembly accuracy in presence of repeats
- Provides extensive assembly information to assist in trouble-shooting assembly problems
- Handles large data sets

Consed

- Automatically chooses finishing reads
- Speeds up finishing
- Integrated with Phrap
- DNA editing more efficient



Sequence from your cDNA clone

TACAGCGCGTCCCCCTCCGGCGTCGCGCTTCTCGATTCCAAGGGGAATGTTTTT
AAAGGCTCTTACATTGAGTCCGCTGCTTATAACCCCAGCTTGGGACCGCTTCA
GGCCGCCATCGTCGCCTTCATCGCCGGCGGCGGTGGGGATTATGAAGAGATTG
TTGCGGCGGTGTTGGTGGAGAAGGAAGGGGCGGTCATCAAACAGGATCACAC
TGCAAGGTTGCTGCTCCATTCCATAGCGCCACGCTGCCACTTCAACAATTTTCT
TGCTTCTCAATCTC

DNA sequence assembly

- PHRAP
- ConSED
- CAP3
- DNASTAR by Lasergene
  - Commercial - http://www.dnastar.com/
- Sequencher- commercial automated sequencers
  - -http://www.genecodes.com/
  - Sequencher protocol
  - http://bip.weizmann.ac.il/sequencher/sequencher.html

cystein protease

GGAGCTCCACCGCGGTGGCGGCCGTTCTAGAACTAGTGGATCCCCCGGGCTGCAGGAATTCGGCACCAGAACAGTGG
GAG
GGAGATCCAAAAAGAGAGAGTGGAAAAGATGGCGCGGTTGATCACAGTGGTGTGGTGCGCGGTGGCGGTGCTATTATG
CG
CCGCGGCCGGTGCGTCGTGGGTGGAGGAGGCGGCGAACCCGATACGAATGGTGTCTGGCGTGGAGGCGGAGGTGGTTC
GG
GTGATCGGGGAGTGCCGGCGTGCGTTGAAGTTTGCTAGGTTCGTGAGCAGGTTCGGGAAGAGTTACCAAAGCGAGGAA
GA
GATGAAGAGAGGTACGAGATATTCTCGCAGAATCTCAGGTTCATCCGCTCCCACAACAAGAAGCGTTTGCCCTATACTC
T
CTCTGTTAATCATTGTGCTGATTGGACTTGGGAGGAGTTCAAAAGACACAGACTAGGAGCTGCCCAAAATTGCTCTGCC
A
CTCTTAACGGCAACCACAAGCTCACCGATGCTGTTCTTCCTCCAACGAAAGACTGGAGAAAAGAAGGTATAGTGAGTT
CA
GTTAAAGATCAAGGCAGCTGCGGATCATGCTGGACATTCAGCACAACTGGGGCTTTAGAAGCAGCCTATGCACAAGCA
TT
TGGGAAGAGTATCTCTCTTTCTGAGCAGCAGCTAGTGGACTGTGCTGGCCCTTTCAACAACTTTGGCTGCCATGGTGGG
T
TGCCATCACAAGCCTTTGAGTACATTAAATACAATGGTGGACTAGAGACAGAGGAAGCATATCCCTACACAGGAAAAG
AT
GGTGTCTGCAAATTCTCAGCTGAAAATGTTGCTGTTCAAGTCCTTGACTCTGTGAATATCACCTTGGGTGCTGAAGATG
A
ACTAAAACATGCAGTTGCATTTGTTCGGCCAGTTAGTGTGGCCTTTCAGGTGGTGAATGGGTTCCATTTCTACGAGAAT
G
GAGTTTTCACTAGTGACACTTGTGGTAGCACTTCCCAGGATGTGAACCATGCCGTTCTTGCTGTTGGATATGGAGTTGA
A
AATGGTGTCCCATATTGGCTCATAAAAAATTCATGGGGAGAAAGCTGGGGTGAAAATGGCTACTTCAAGATGGAATTG
GG
GAAGAACATGTGTGGTGTTGCAACTTGTGCATCTTATCCAATTGTGGCATAAATTGCATAACAATATGGTCCCTGGTGA
C
TACCACCTTGTCATGCTTCAGAGTTTAGAGCGTATTGCTGATGCCAGTATTGATGAATGATGATTAAAGATAAGTGTAA
T
TGTATGATGAAAATTGCTCCTAGTTGGGTTGGCATGATGTATAAAATAAGCTAGAAGTTGTTGTAAATCACATAAGTAT
A
TTATGGCCTTAATTGTGTGTATCACAGACATAATAACGATCATATATTGATAGTTCATAGTTACATATTGTATTGTATTG
ATGCTCCCGCTTCAAATATCAGTTATAAGATAGCATTGTCTTTGCTACTTTGCACTATGCAACATTATT

Sequence Alignments

Why do DNA sequence alignments?

- If your sequence is not full length, then add other expressed sequence tags (ESTs) to build full-length clone
- Can identify mismatches for single nucleotide polymorphism (SNP) discovery
- Provide a measure of relatedness between nucleotide sequences
- Usually <u>protein alignments</u> with other proteins are used to determining relatedness that allows the drawing of biological inferences regarding
  - Structural relationships
  - Functional relationships
  - Evolutionary relationships

Similarity

- A quantitative measure
- Based on an observable
- Usually expressed as percent identity
- Quantifies changes that occur as two sequences diverge
    - Substitutions
    - Insertions
    - Deletions
  - Identifies residues crucial for maintaining a protein's structure or function

Similarity

- High degrees of similarity *might* imply
  - A common evolutionary history
  - A possible commonality in biological function

Homology

- Implies an evolutionary relationship
- Hay apply to the relationship
  - Between genes separated by the event of speciation (orthology), ie. orthologous genes
  - Between genes separated by the event of genetic duplication (paralogy), ie. paralogous genes

- Orthologs
  - Sequences are direct descendants of a sequence in a common ancestor
  - Most likely have similar domain structure, three dimensional structure, and biological function
- Paralogs
  - Related through a gene duplication event
  - Provides insight into evolution, ie. adapting a pre-existing gene product for a new function

Global Sequence Alignments

- Sequence comparison along the entire length of two sequences being aligned
- Best for highly similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss relationships

Local Sequence Alignments

- Sequence comparison intended to find the <u>most similar regions</u> in two sequences being aligned
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated
- Best for sequences that share some similarity or for <u>sequences of different lengths</u>

Scoring Matrices

- Empirical weighting scheme to represent biology
- DNA only has A,T,G,C
- Protein has amino acids; relatedness among amino acids; function; charges; side groups

Matrix Structure: Nucleotides

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-1	-3	-1
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-1	-1	-1
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-1	-1	-1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1
B	-4	-1	-1	-1	-1	-3	-1	-1	-1	-3	-1	-2	-2	-2	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-3	-1	-2	-2	-1
H	-1	-1	-4	-1	-1	-1	-1	-1	-3	-1	-3	-2	-1	-2	-1
D	-1	-1	-1	-4	-1	-1	-1	-3	-1	-1	-3	-2	-3	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

- *Simple match/mismatch scoring scheme*
- *Assumes each nucleotide occurs 25% of the time*

Sequence alignment – Building a contiguous sequence "contig"

Building a contig

- ESTs must be from the same gene, not a paralog (gene duplication event)
- ESTs must be of high quality sequence
- After a contig is constructed, the sequence should be confirmed by cloning and sequencing

EST 1: gagcctatgccgtccgagattacgggcttacaggattcagatt

EST 2: ggacccaagtttcacgtccaatattgtgttgaccatagaaaaaaaaa

EST 3: acgggcttacaggattcagattcatggacccaagtttcacgtc

EST alignment to make contig

EST 1: gagcctatgccgtccgagattacgggcttacaggattcagatt

EST 3: acgggcttacaggattcagattcatggacccaagtttcacgtc

EST2: ggacccaagtttcacgtccaatattgtgttgacc atagaaaaaaaaa

Consensus sequence:

gagcctatgcgtccgagattacgggcttacaggattcagattcatggaccaagttcacgtccaatattgtgttgaccata gaaaaaaaaa

In this example EST 3 forms a bridge to connect EST 1 and EST 2

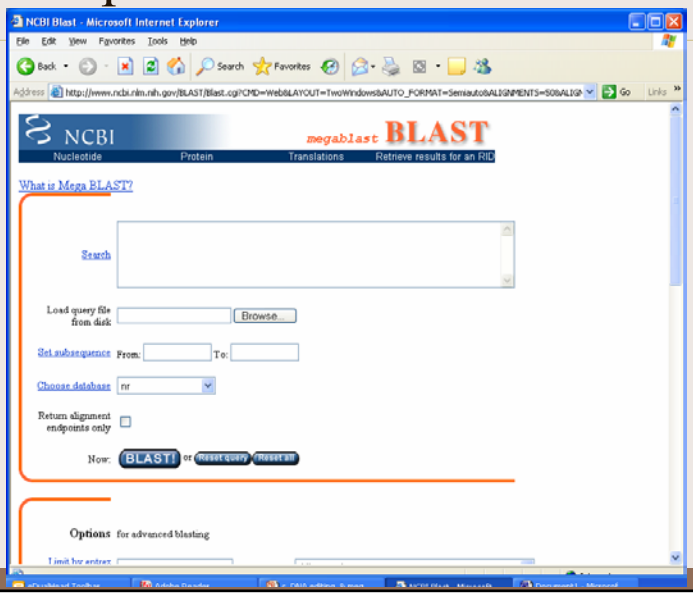
DNA STAR EXAMPLE

Making a contig from EST sequences

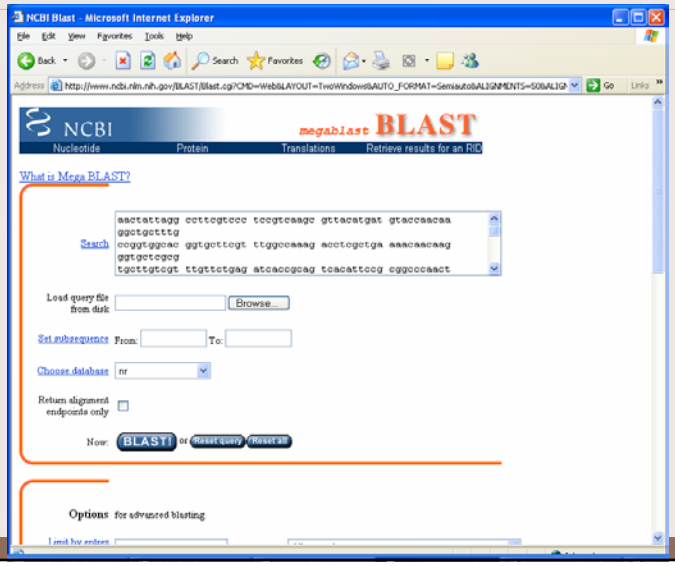
This is your sequence from a clone

1 aactattagg ccttcgtccc tccgtcaagc gttacatgat gtaccaacaa ggctgctttg
61 ccggtggcac ggtgcttcgt ttggccaaag acctcgctga aaacaacaag ggtgctcgcg
121 tgcttgtcgt ttgttctgag atcaccgcag tcacattccg cggcccaact gacacccatc
181 ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc agccgctgtc attgttggat
241 cagacccctt accagttgaa aagcctttgt ttcagcttat ctggactgcc caaacaatcc
301 ttccagacag tgaaggggct attgatggcc accttcgcga agttggactc actttccatc
361 tcctcaagga tgttcctgga ctcatctcta agaatattga gaaggccttg gttgaagcct
421 tccaaccctt gggaatctcc gattacaatt ctatcttctg gattgcacac cct

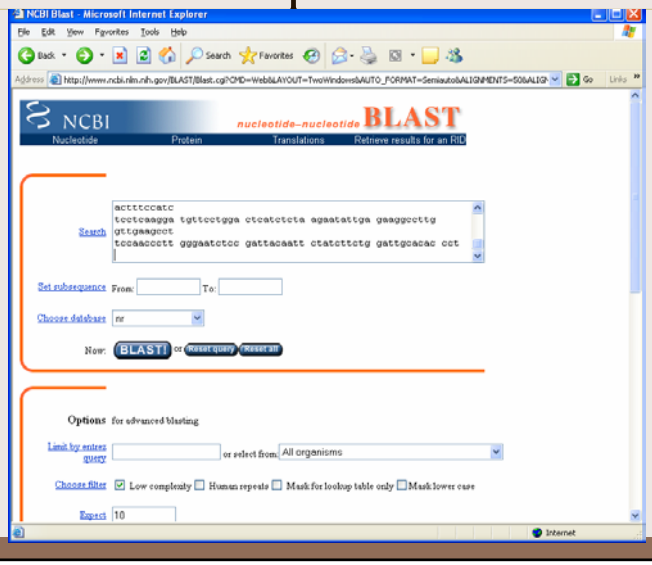
Step 1:Go to BLAST search

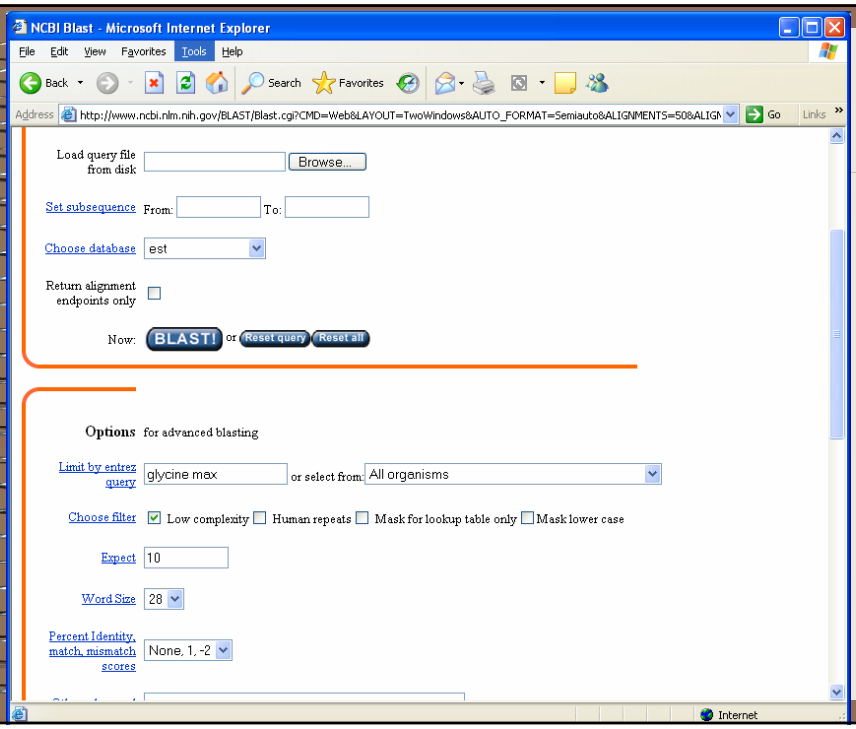


Step 2: Paste sequence

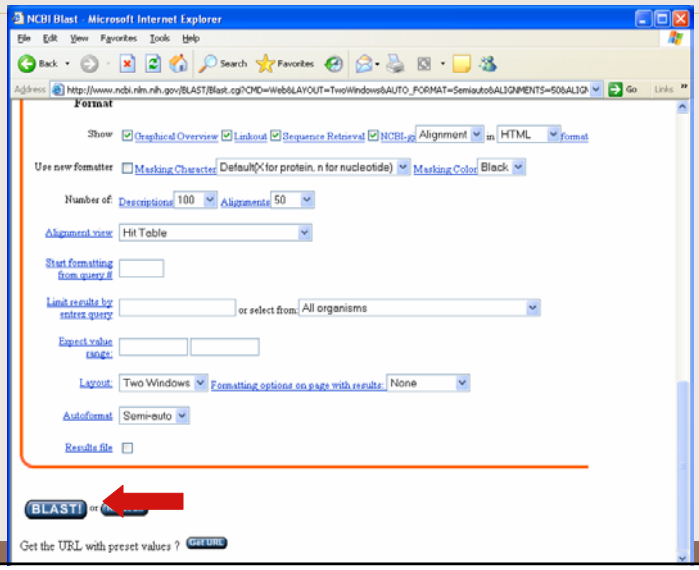


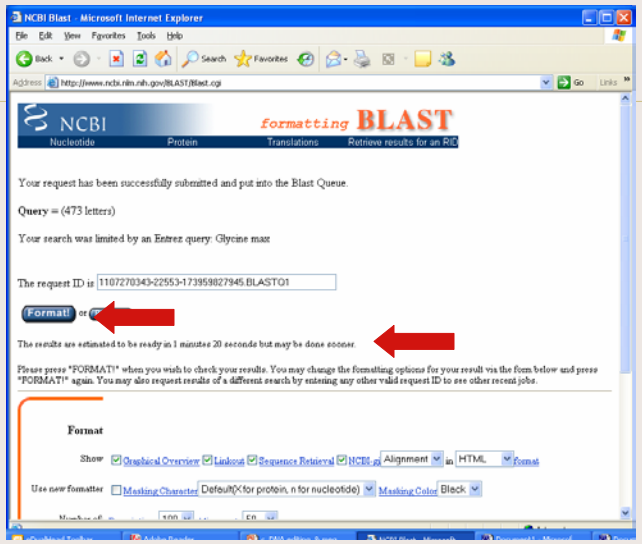
Step 3: Set constraints and options

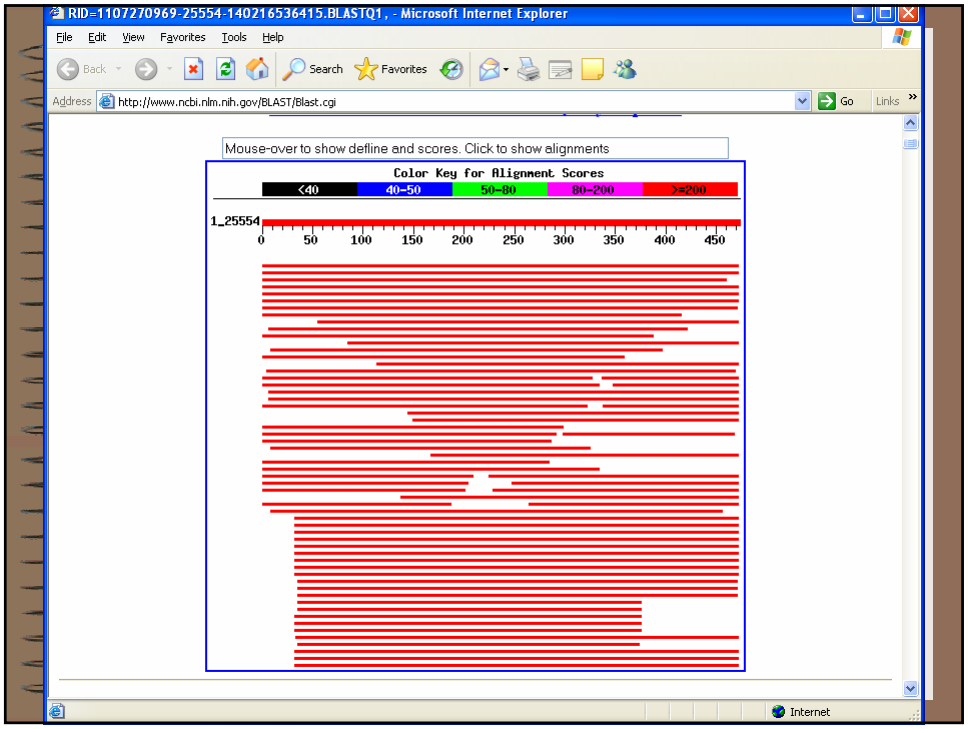


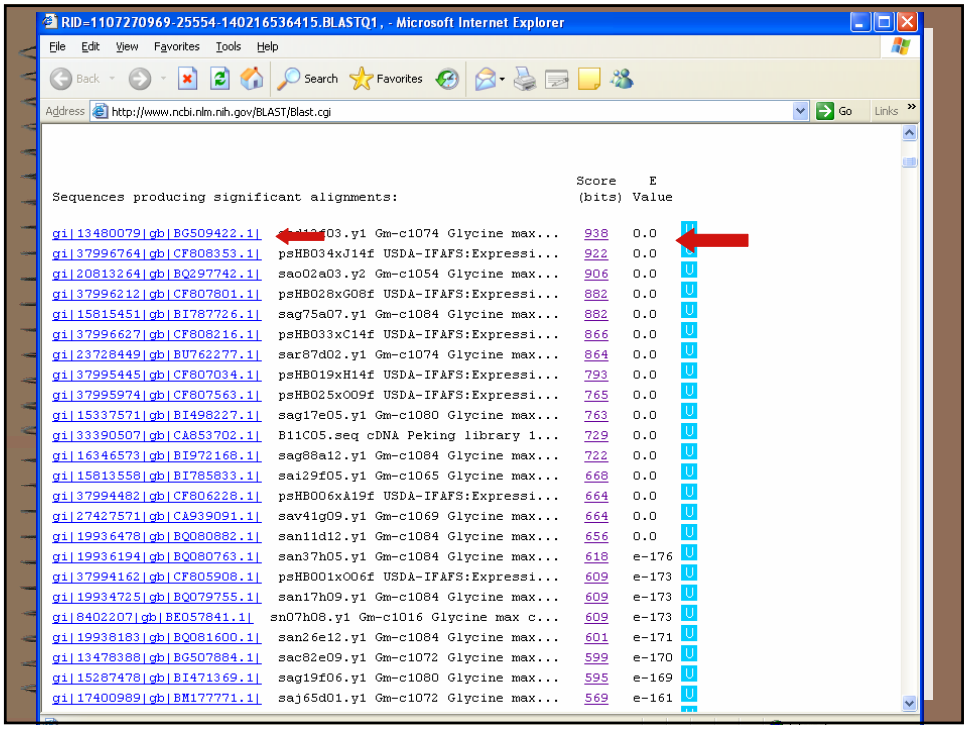


Step 4: BLAST!


NCBI Blast - Microsoft Internet Explorer
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
Alignment
in
HTML
format
Use new formatter
Masking Character
Default(X for protein, n for nucleotide)
Masking Color
Black
Number of:
Descriptions
100
Alignments
50
Alignment view
Hit Table
Start formatting from query #
Limit results by entrez query
or select from
All organisms
Expect value range:
Layout:
Two Windows
Formatting options on page with results:
None
Autoformat
Semi-auto
Results file
BLAST!
Get the URL with preset values ?
Get URL


NCBI Blast - Microsoft Internet Explorer
NCBI
formatting BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = (473 letters)
Your search was limited by an Entrez query: Glycine max
The request ID is
1107270343-22553-173959827945.BLASTQ1
Format!
The results are estimated to be ready in 1 minutes 20 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
Alignment
in
HTML
format
Use new formatter
Masking Character
Default(X for protein, n for nucleotide)
Masking Color
Black


RID=1107270969-25554-140216536415.BLASTQ1 , - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40 40-50 50-80 80-200 >=200
1_25554
0 50 100 150 200 250 300 350 400 450
Internet


RID=1107270969-25554-140216536415.BLASTQ1 , - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Sequences producing significant alignments: Score (bits) E Value
gi|13480079|gb|BG509422.1| 938 0.0
gi|37996764|gb|CF808353.1| psHB034xJ14f USDA-IFAFS:Expressi... 922 0.0
gi|20813264|gb|BQ297742.1| sao02a03.y2 Gm-c1054 Glycine max... 906 0.0
gi|37996212|gb|CF807801.1| psHB028xG08f USDA-IFAFS:Expressi... 882 0.0
gi|15815451|gb|BI787726.1| sag75a07.y1 Gm-c1084 Glycine max... 882 0.0
gi|37996627|gb|CF808216.1| psHB033xC14f USDA-IFAFS:Expressi... 866 0.0
gi|23728449|gb|BU762277.1| sar87d02.y1 Gm-c1074 Glycine max... 864 0.0
gi|37995445|gb|CF807034.1| psHB019xH14f USDA-IFAFS:Expressi... 793 0.0
gi|37995974|gb|CF807563.1| psHB025xO09f USDA-IFAFS:Expressi... 765 0.0
gi|15337571|gb|BI498227.1| sag17e05.y1 Gm-c1080 Glycine max... 763 0.0
gi|33390507|gb|CA853702.1| B11C05.seq cDNA Peking library 1... 729 0.0
gi|16346573|gb|BI972168.1| sag88a12.y1 Gm-c1084 Glycine max... 722 0.0
gi|15813558|gb|BI785833.1| sai29f05.y1 Gm-c1065 Glycine max... 668 0.0
gi|37994482|gb|CF806228.1| psHB006xA19f USDA-IFAFS:Expressi... 664 0.0
gi|27427571|gb|CA939091.1| sav41g09.y1 Gm-c1069 Glycine max... 664 0.0
gi|19936478|gb|BQ080882.1| san11d12.y1 Gm-c1084 Glycine max... 656 0.0
gi|19936194|gb|BQ080763.1| san37h05.y1 Gm-c1084 Glycine max... 618 e-176
gi|37994162|gb|CF805908.1| psHB001xO06f USDA-IFAFS:Expressi... 609 e-173
gi|19934725|gb|BQ079755.1| san17h09.y1 Gm-c1084 Glycine max... 609 e-173
gi|8402207|gb|BE057841.1| sn07h08.y1 Gm-c1016 Glycine max c... 609 e-173
gi|19938183|gb|BQ081600.1| san26e12.y1 Gm-c1084 Glycine max... 601 e-171
gi|13478388|gb|BG507884.1| sac82e09.y1 Gm-c1072 Glycine max... 599 e-170
gi|15287478|gb|BI471369.1| sag19f06.y1 Gm-c1080 Glycine max... 595 e-169
gi|17400989|gb|BM177771.1| saj65d01.y1 Gm-c1072 Glycine max... 569 e-161

NCBI Sequence Viewer v2.0 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=13480079&dopt=GenBank Go Links

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Send all to file

Range: from begin to end Reverse complemented strand Features: SNP CDD MGC HPRD

1: BG509422. Reports sad13f03.y1 Gm-c1...[gi:13480079] Links

```
LOCUS       BG509422                 473 bp    mRNA    linear   EST 24-JUL-2004
DEFINITION  sad13f03.y1 Gm-c1074 Glycine max cDNA clone GENOME SYSTEMS CLONE
            ID: Gm-c1074-246 5' similar to SW:CHS1_SOYBN P24826 CHALCONE
            SYNTHASE 1 ;, mRNA sequence.
ACCESSION   BG509422
VERSION     BG509422.1  GI:13480079
KEYWORDS    EST.
SOURCE      Glycine max (soybean)
  ORGANISM  Glycine max
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids
            ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae;
            Glycine.
REFERENCE   1  (bases 1 to 473)
  AUTHORS   Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna
            ,A., Bolla,B., Marra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C.,
            Wylie,T., Underwood,K., Steptoe,M., Theising,B., Allen,M., Bowers
            ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk
            ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann
            ,R., Waterston,R. and Wilson,R.
  TITLE     Public Soybean EST Project
  JOURNAL   Unpublished (1999)
```

Internet

NCBI Sequence Viewer v2.0 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=13480079&dopt=GenBank Go Links

```
            tissue with Pseudomonas syringae pv. glycinea carrying the
            avrB gene (Genetics 141:1597-1604). Plant tissue (expanded
            unifoliate leaves) was collected at 2, 4, 8, 12, 24, 36,
            and 53 hrs after inoculation and their mRNA pooled equally
            for cDNA construction. The  library was prepared using the
            Stratagene pBluescript II SK(+) library construction kit.
            Complementary DNA was synthesized from mRNA using a primer
            consisting of a poly(dT) sequence with an XhoI restriction
            site. EcoRI adaptors were ligated to the blunt-ended cDNA
            fragments followed by XhoI digestion. The cDNA insert is
            protected from XhoI digestion via methylation during first
            strand synthesis. The cDNA fragments were directionally
            cloned into the EcoRI-XhoI restriction site of the
            pBluescript vector. The ligated cDNA fragments were
            transformed into E.coli ElectroMax DH10B host cells. Plant
            care, inoculations, and library construction were
            performed by Steve Clough (Lila Vodkin lab, University of
            Illinois)."
ORIGIN
        1 aactattagg ccttcgtccc tccgtcaagc gttacatgat gtaccaacaa ggctgctttg
       61 ccggtggcac ggtgcttcgt ttggccaaag acctcgctga aaacaacaag ggtgctcgcg
      121 tgcttgtcgt ttgttctgag atcaccgcag tcacattccg cggcccaact gacacccatc
      181 ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc agccgctgtc attgttggat
      241 cagacccctt accagttgaa aagccttttgt ttcagcttat ctggactgcc caaacaatcc
      301 ttccagacag tgaaggggct attgatggcc accttcgcga agttggactc actttccatc
      361 tcctcaagga tgttcctgga ctcatctcta agaatattga gaaggccttg gttgaagcct
      421 tccaaccctt gggaatctcc gattacaatt ctatcttctg gattgcacac cct
//
```

Disclaimer | Write to the Help Desk

NCBI | NLM | NIH

Jan 27 2005 17:14:21

Internet

Collect sequences into a series of files so they can be aligned

File 1.

1 aactattagg ccttcgtccc tccgtcaagc gttacatgat gtaccaacaa ggctgctttg
61 ccggtggcac ggtgcttcgt ttggccaaag acctcgctga aaacaacaag ggtgctcgcg
121 tgcttgtcgt ttgttctgag atcaccgcag tcacattccg cggcccaact gacacccatc
181 ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc agccgctgtc attgttggat
241 cagacccctt accagttgaa aagcctttgt ttcagcttat ctggactgcc caaacaatcc
301 ttccagacag tgaaggggct attgatggcc accttcgcga agttggactc actttccatc
361 tcctcaagga tgttcctgga ctcatctcta agaatattga gaaggccttg gttgaagcct
421 tccaaccctt gggaatctcc gattacaatt ctatcttctg gattgcacac cct

File 2

1 ggcaatcaag gaatggggtc aacccaagtc caagattacc catctcatct tttgcaccac
61 tagtggtgtc gacatgcctg gtgctgatta tcagctcact aaactattag gccttcgtcc
121 ctccgtcaag cgttacatga tgtaccaaca aggctgcttt gccggtggca cggtgcttcg
181 tttggccaaa gacctcgctg aaaacaacaa gggtgctcgc gtgcttgtcg tttgttctga
241 gatcaccgca gtcacattcc gcggcccaat tgacacccat cttgatagcc ttgtgggtca
301 agccttgttt ggagatggtg cagccgctgt cattgttgga tcagacccct taccagttga
361 aaagcctttg tttcagctta tctggactgc ccaaacaatc cttccagaca gtgaaggggc
421 tattgatggc caccttcgcg aagttggact cactttccat ctcctcaagg atgttcctgg
481 actcatctct aagaatattg agaaggcttt ggttgaagcc ttccaaccct tgggaatctc
541 cgattacaat tctatcttct ggattgcaca ccctggtgga cccgcaattt tggaccaagt
601 tgaggctaag ttaggcttga agcctgaaaa aatggaagct actagacatg tgctcagcga
661 gtatggtaac atgt

File 3

1 tgtggaggta ccaaagttgg gaaaagaggc tgcaactaag gcaatcaagg aatggggtca
61 acccaagtcc aagattaccc atctcatctt ttgcaccact agtggtgtcg acatgcctgg
121 tgctgattat cagctcacta aactattagg ccttcgtccc tccgtcaagc gttacatgat
181 gtaccaacaa ggctgctttg ccggtggcac ggtgcttcgt ttggccaaag acctcgctga
241 aaacaacaag ggtgctcgcg tgcttgtcgt ttgttctgag atcaccgcag tcacattccg
301 cggcccaact gacacccatc ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc
361 agccgctgtc attgttggat cagacccctt accagttgaa aagcctttgt ttcagcttat
421 ctggactgcc caaacaatcc ttccagacag tgaaggggct attgatggcc accttcgcga
481 agttggactc actttccatc tcctcaagga tgttcctgga ctcatctcta agaatattga
541 gaaggccttg gttgaagcct tcccaccctt gggaatctcc gattacaatt ctatcttctg
601 g

DNA STAR

- Edit sequence
  - Allows you to import and edit DNA and protein sequences
- Megalign
  - Allows you to align DNA and protein sequences

Edit Sequence

EditSeq

File Edit Search Speech Features Goodies Windows Net Search Help

Untitled Seq #1 : SEQUENCE

Position: 471 470 bp

```
10        20        30        40        50        60
aactattaggccttcgtccctccgtcaagcgttacatgatgtaccaacaaggctgctttg      60
ccggtggcacggtgcttcgtttggccaaagacctcgctgaaaacaacaagggtgctcgcg     120
tgcttgtcgtttgttctgagatcaccgcagtcacattccgcggcccaactgacacccatc     180
ttgatagccttgtgggtcaagccttgtttggagatggtgcagccgctgtcattgttggat     240
cagacccccttaccagttgaaaagcctttgtttcagcttatctggactgcccaaacaatcc     300
ttccagacagtgaaggggctattgatggccaccttcgcgaagttggactcactttccatc     360
tcctcaaggatgttcctggactcatctctaagaatattgagaaggccttggttgaagcct     420
tcccaacccttgggaatctccgattacaattctatcttctggattgcacac              470
```

Created: Tuesday, February 01, 2005 10:45 AM

Unspecified Search

EditSeq

File Edit Search Speech Features Goodies Windows Net Search Help

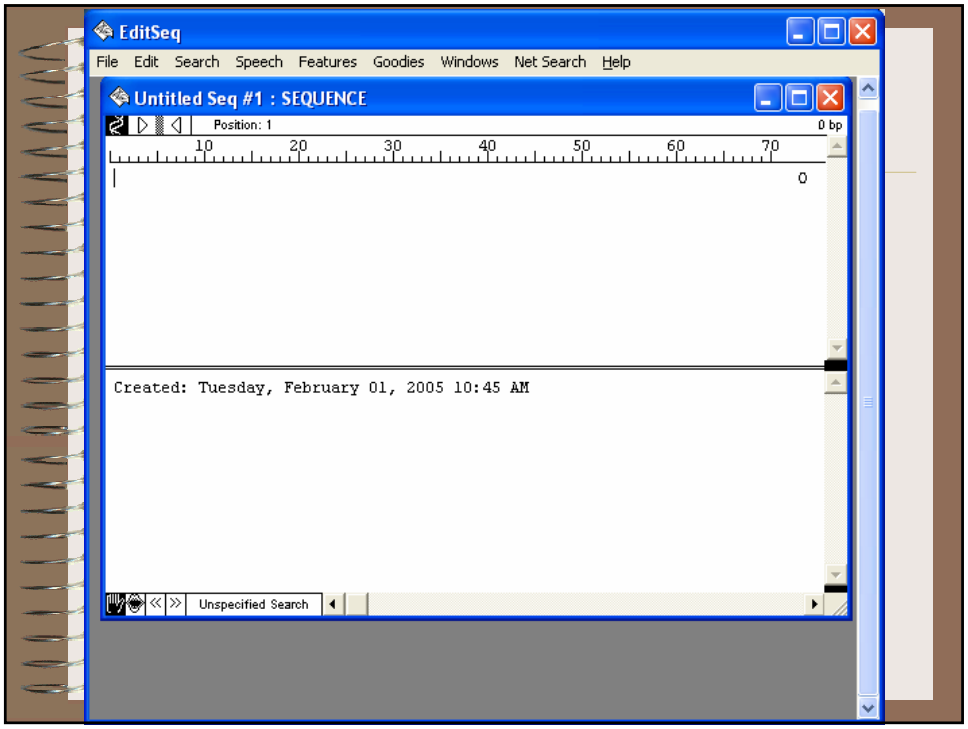
soybean1.seq : SEQUENCE

Position: 1 473 bp

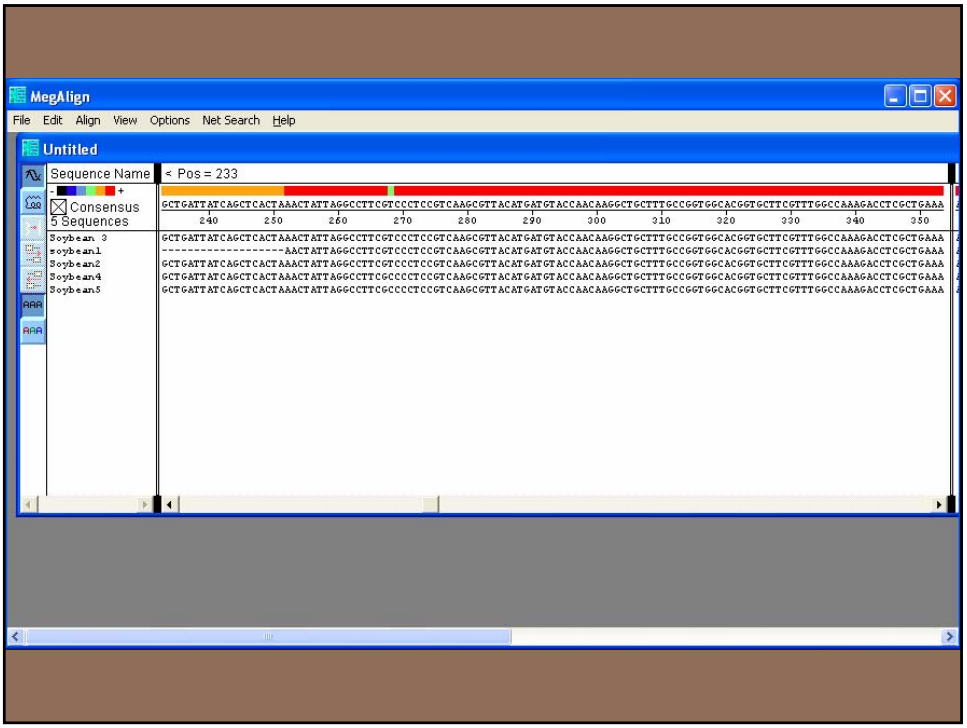
```
10        20        30        40        50        60
aactattaggccttcgtccctccgtcaagcgttacatgatgtaccaacaaggctgctttg      60
ccggtggcacggtgcttcgtttggccaaagacctcgctgaaaacaacaagggtgctcgcg     120
tgcttgtcgtttgttctgagatcaccgcagtcacattccgcggcccaactgacacccatc     180
ttgatagccttgtgggtcaagccttgtttggagatggtgcagccgctgtcattgttggat     240
cagacccccttaccagttgaaaagcctttgtttcagcttatctggactgcccaaacaatcc     300
ttccagacagtgaaggggctattgatggccaccttcgcgaagttggactcactttccatc     360
tcctcaaggatgttcctggactcatctctaagaatattgagaaggccttggttgaagcct     420
tcccaacccttgggaatctccgattacaattctatcttctggattgcacaccct           473
```

Created: Tuesday, February 01, 2005 10:39 AM

Unspecified Search


EditSeq
File Edit Search Speech Features Goodies Windows Net Search Help
Untitled Seq #1 : SEQUENCE
Position: 1
0 bp
10 20 30 40 50 60 70
0
Created: Tuesday, February 01, 2005 10:45 AM
Unspecified Search

Megalign



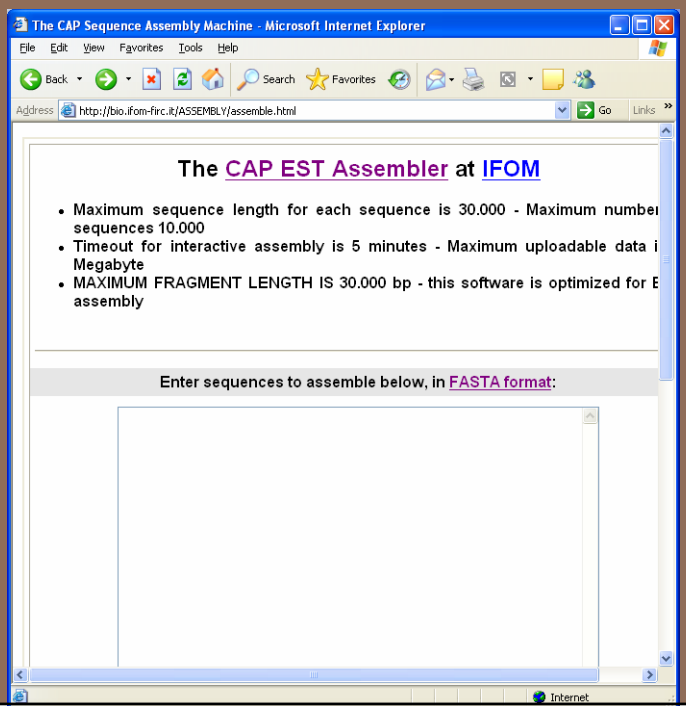
CAP EST Assembler
Contig Assembly Program

- http://bio.ifom-firc.it/ASSEMBLY/assemble.html
- Can use up to 30,000 EST sequences
- Fragment maximum is 30,000 bp
- Sequences must be in FASTA format
- Huang, X. 1992. Genomics 14: 18-25
- Huang, X. 1996. Genomics 33: 21-31

Edit file

- **Some** DNA and protein alignment software requires a specific format
- FASTA format
  - Header HAS TO start with ‘ > ’
  - A description should follow
  - For DNA only five letters A,C,T,G,N allowed
  - No numbers

```
>soybean1
ATTCCTTAGGATC…
>soybean 2
TCCGTCAGGTGTT…
>soybean3
GGCTATGGCCTAAT…
```



The CAP Sequence Assembly Machine - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://bio.ifom-firc.it/ASSEMBLY/assemble.html Go Links

- Maximum sequence length for each sequence is 30.000 - Maximum number of sequences 10.000
- Timeout for interactive assembly is 5 minutes - Maximum uploadable data is 1 Megabyte
- MAXIMUM FRAGMENT LENGTH IS 30.000 bp - this software is optimized for EST assembly

---

Enter sequences to assemble below, in FASTA format:

```
>Soybean1
aatggggtcaacccaagtccaagattacccatctcatcttttgcaccactagtggtgtc
gacatgcctggtgctgattat
cagctcactaaactattaggccttcgcccctccgtcaagcgttacatgatgtaccaaca
aggctgctttgccggtggcac
ggtgcttcgtttggccaaagacctcgctgaaaacaacaagggtgctcgcgtgcttgtcg
tttgttctgagatcaccgcag
tcacatttcgcggcccaactgacacccatcttgatagccttgtgggtcaagccttgttt
ggagatggtgcagccgctgtc
attgttggatcagacccctttaccagttgaaaagcctttgtttcagcttgtctggactgc
ccagacaatccttccagacag
tgaaggggctattgatggacaccttcgcgaagttggtctcactttccatctcctcaagg
atgttcctggactcatctcca
agaatattgagaaggccttggttgaagccttccaacccttgggaatctccgattacaat
tctatcttctggattgcacac
cct
>Soybean2
ggcaatcaaggaatggggtcaacccaagtccaagattacccatctcatcttttgcacca
ctagtggtgtcgacatgcctg
gtgctgattatcagctcactaaactattaggccttcgtccctccgtcaagcgttacatg
```

Internet

Cap3 Assembler at IFOM output - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://bio.ifom-firc.it/cgi-bin/Assembly/capassemble.pl Go Links

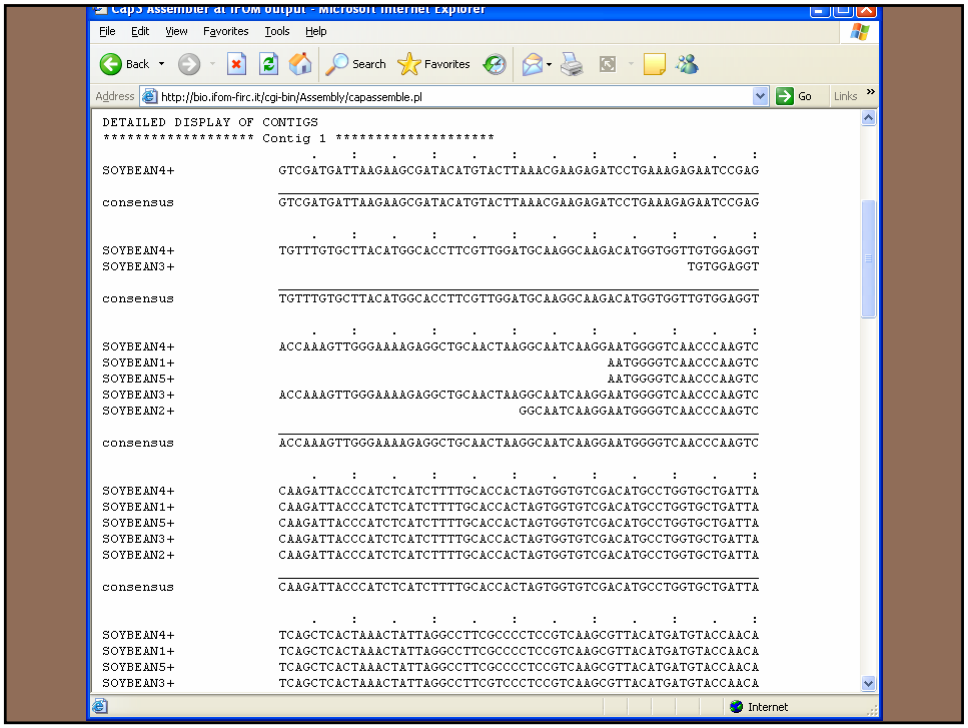
The Cap Sequence Assembler at IFOM

Program Cap3 used

Assembly results:

```
Number of segment pairs = 20; number of pairwise comparisons = 10
'+' means given segment; '-' means reverse complement

Overlaps            Containments  No. of Constraints Supporting Overlap

******************** Contig 1 ********************
SOYBEAN4+
                    SOYBEAN1+ is in SOYBEAN4+
                    SOYBEAN5+ is in SOYBEAN1+
                    SOYBEAN3+ is in SOYBEAN4+
SOYBEAN2+

DETAILED DISPLAY OF CONTIGS
******************** Contig 1 ********************
                          .    :    .    :    .    :    .    :    .    :    .    :
SOYBEAN4+             GTCGATGATTAAGAAGCGATACATGTACTTAAACGAAGAGATCCTGAAAGAGAATCCGAG
                      ------------------------------------------------------------
consensus             GTCGATGATTAAGAAGCGATACATGTACTTAAACGAAGAGATCCTGAAAGAGAATCCGAG

                          .    :    .    :    .    :    .    :    .    :    .    :
SOYBEAN4+             TGTTTGTGCTTACATGGCACCTTCGTTGGATGCAAGGCAAGACATGGTGGTTGTGGAGGT
SOYBEAN3+                                                                TGTGGAGGT
                      ------------------------------------------------------------
consensus             TGTTTGTGCTTACATGGCACCTTCGTTGGATGCAAGGCAAGACATGGTGGTTGTGGAGGT

                          .    :    .    :    .    :    .    :    .    :    .    :
```

Internet



What we learned today

- DNA editing
- Phred
- Phrap
- Consed
- DNA Sequencing software
- DNA sequence assembly
- Similarity searching with a DNA sequence
- BLAST