Sequence Similarity Searching

Why Compare Sequences?

- Identify sequences found in lab experiments
  - What is this thing I just found?
- Compare new genes to known ones
- Compare genes from different species
  - information about evolution
- Guess functions for entire genomes full of new gene sequences

Are there other sequences like this one?

1) Huge public databases - GenBank, Swissprot, etc.

2) Sequence comparison is the most powerful and reliable method to determine evolutionary relationships between genes

3) Similarity searching is based on alignment

4) **BLAST** and **FASTA** provide rapid similarity searching

   a. rapid = approximate (heuristic)

   b. false + and - scores

Similarity is based on Alignment

```
GATGCCATAGAGCTGTAGTCGTACCCT  <-
->  CTAGAGAGC-GTAGTCAGAGTGTCTTTGAGTTCC
```

Similarity ≠ Homology

1) 25% similarity ≥ 100 AAs is strong evidence for homology

2) Homology is an evolutionary statement which means “descent from a common ancestor”

   – common 3D structure

   – usually common function

   – homology is all or nothing, you cannot say "50% homologous"

Alignment is Based on Dot Plots

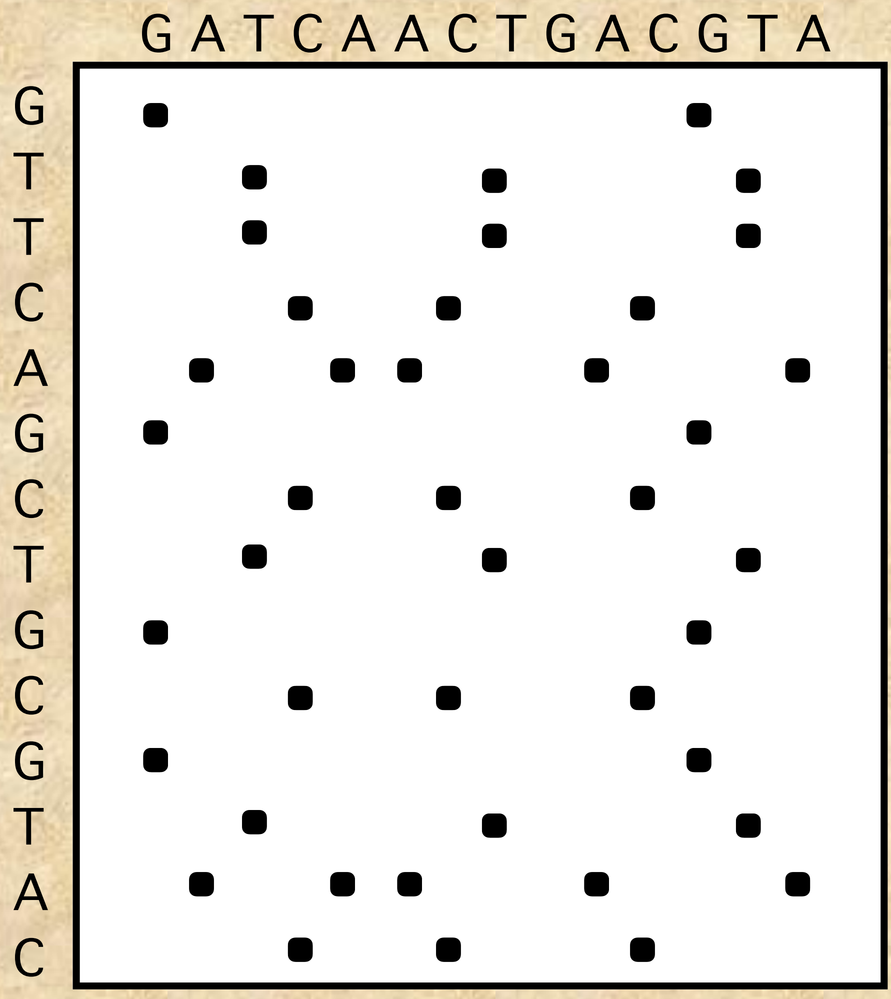
1) two sequences on vertical and horizontal axes of graph

2) put dots wherever there is a match

3) diagonal line is region of identity
(local alignment)

4) apply a window filter - look at a group of bases, must meet % identity to get a dot

Simple Dot Plot

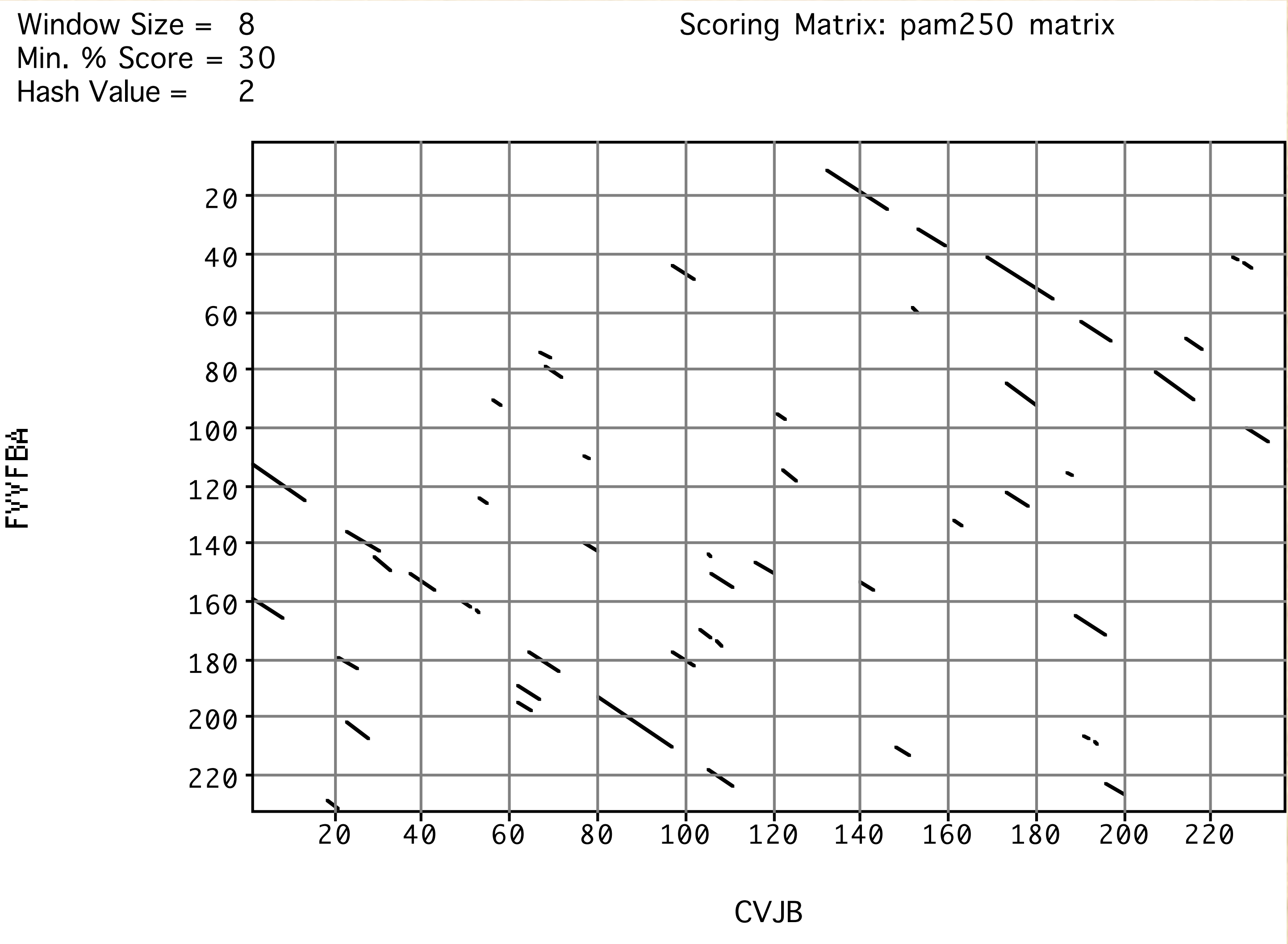


	G	A	T	C	A	A	C	T	G	A	C	G	T	A
G	●											●		
T			●					●					●	
T			●					●					●	
C				●			●				●			
A		●			●	●				●				●
G	●											●		
C				●			●				●			
T			●					●					●	
G	●											●		
C				●			●				●			
G	●											●		
T			●					●					●	
A		●			●	●				●				●
C				●			●				●			

Dot plot filtered with 4 base window and 75% identity

	G	A	T	C	A	A	C	T	G	A	C	G	T	A
G	●													
T		●												
T			●											
C				●										
A					●									
G						●								
C							●							
T														
G									●					
C										●				
G											●			
T														
A														
C														

Dot plot of real data



Window Size = 8
Min. % Score = 30
Hash Value = 2

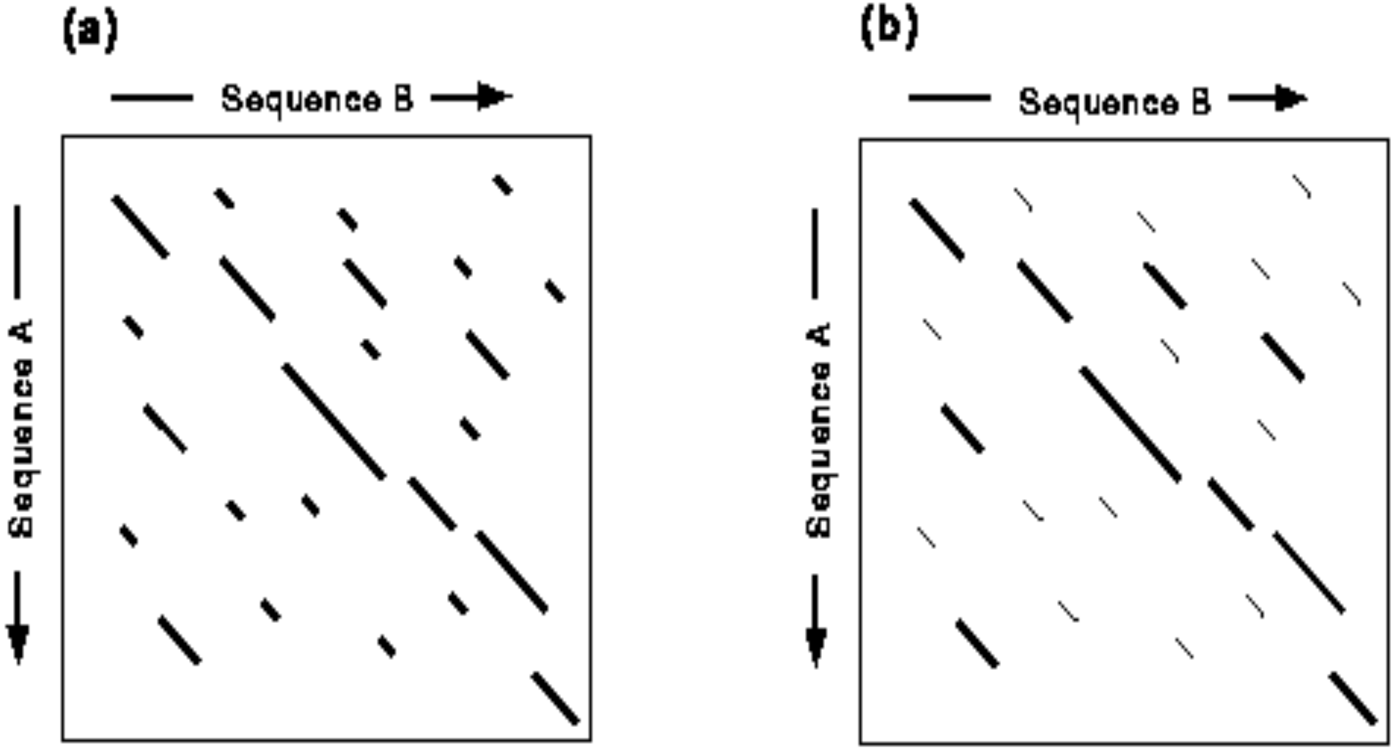
Scoring Matrix: pam250 matrix

FASTA

1) Derived from logic of the dot plot
   - compute best diagonals from all frames of alignment

2) Word method looks for exact matches between words in query and test sequence
   - hash tables (fast computer technique)
   - DNA words are usually 6 bases
   - protein words are 1 or 2 amino acids
   - only searches for diagonals in region of word matches = faster searching

FASTA Algorithm



Find runs of identitical words

Re-score using PAM matrix

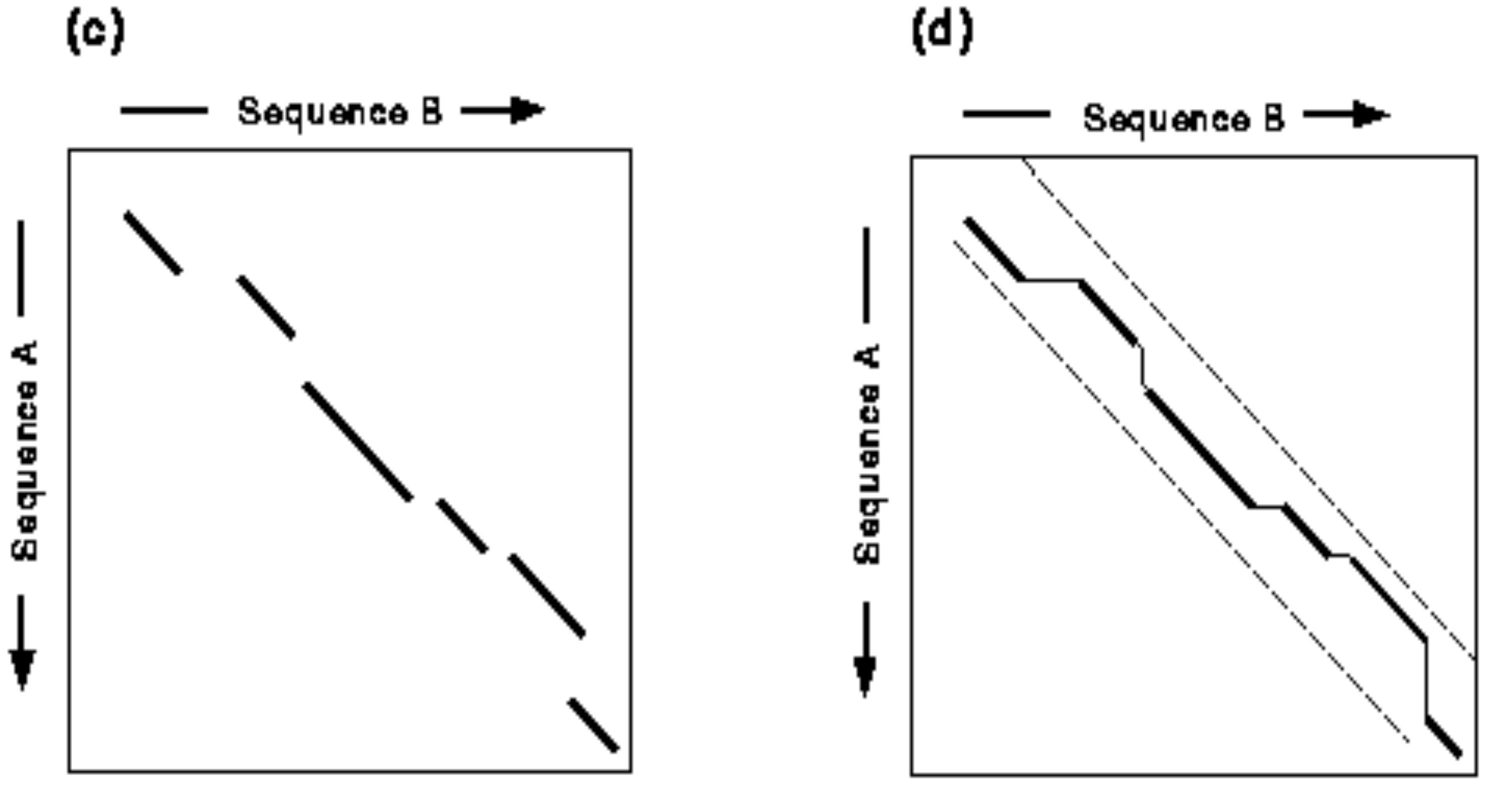
Keep top scoring segments

Makes Longest Diagonal

3) after all diagonals found, tries to join diagonals by adding gaps

4) computes alignments in regions of best diagonals

FASTA Alignments



Join segments using gaps, eliminate other segments

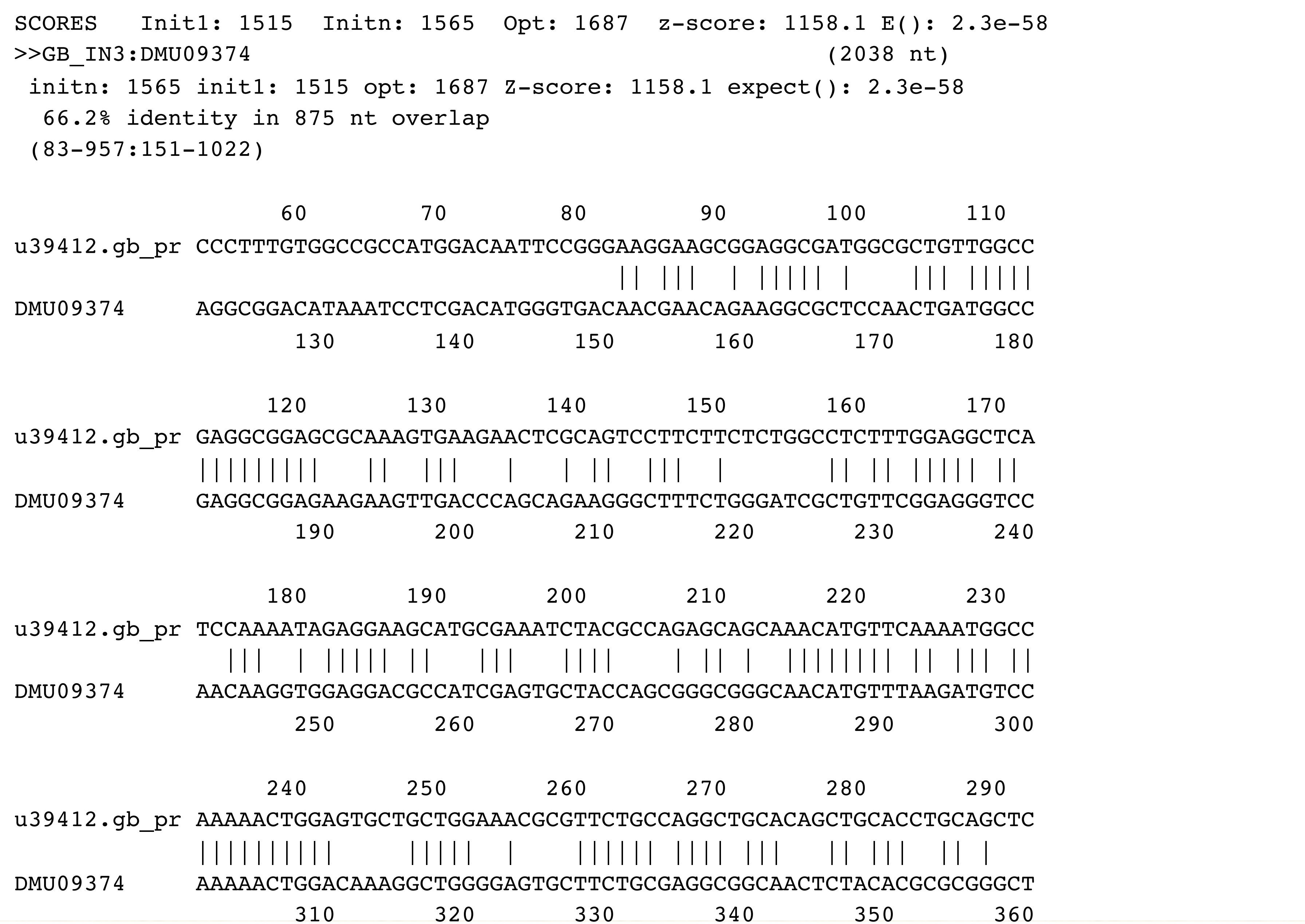
Use dynamic programming to create an optimal alignment

FASTA Results - List

```
The best scores are:                    init1 initn   opt    z-sc E(1018780)..

SW:PPI1_HUMAN    Begin: 1  End: 269
! Q00169 homo sapiens (human). phosph... 1854  1854  1854  2249.3  1.8e-117
SW:PPI1_RABIT    Begin: 1  End: 269
! P48738 oryctolagus cuniculus (rabbi... 1840  1840  1840  2232.4  1.6e-116
SW:PPI1_RAT    Begin: 1  End: 270
! P16446 rattus norvegicus (rat). pho... 1543  1543  1837  2228.7  2.5e-116
SW:PPI1_MOUSE    Begin: 1  End: 270
! P53810 mus musculus (mouse). phosph... 1542  1542  1836  2227.5  2.9e-116
SW:PPI2_HUMAN    Begin: 1  End: 270
! P48739 homo sapiens (human). phosph... 1533  1533  1533  1861.0  7.7e-96
SPTREMBL_NEW:BAC25830    Begin: 1  End: 270
! Bac25830 mus musculus (mouse). 10, ... 1488  1488  1522  1847.6  4.2e-95
SP_TREMBL:Q8N5W1    Begin: 1  End: 268
! Q8n5w1 homo sapiens (human). simila... 1477  1477  1522  1847.6  4.3e-95
SW:PPI2_RAT    Begin: 1  End: 269
! P53812 rattus norvegicus (rat). pho... 1482  1482  1516  1840.4  1.1e-94
```

FASTA Results - Alignment



```
SCORES   Init1: 1515  Initn: 1565  Opt: 1687  z-score: 1158.1 E(): 2.3e-58
>>GB_IN3:DMU09374                                         (2038 nt)
 initn: 1565 init1: 1515 opt: 1687 Z-score: 1158.1 expect(): 2.3e-58
  66.2% identity in 875 nt overlap
 (83-957:151-1022)

                   60        70        80        90       100       110
u39412.gb_pr CCCTTTGTGGCCGCCATGGACAATTCCGGGAAGGAAGCGGAGGCGATGGCGCTGTTGGCC
                                           || |||  | ||||| |    ||| |||||
DMU09374     AGGCGGACATAAATCCTCGACATGGGTGACAACGAACAGAAGGCGCTCCAACTGATGGCC
                    130       140       150       160       170       180

                  120       130       140       150       160       170
u39412.gb_pr GAGGCGGAGCGCAAAGTGAAGAACTCGCAGTCCTTCTTCTCTGGCCTCTTTGGAGGCTCA
             |||||||||   ||  |||   |   | ||  |||  |       || || ||||| ||
DMU09374     GAGGCGGAGAAGAAGTTGACCCAGCAGAAGGGCTTTCTGGGATCGCTGTTCGGAGGGTCC
                    190       200       210       220       230       240

                  180       190       200       210       220       230
u39412.gb_pr TCCAAAATAGAGGAAGCATGCGAAATCTACGCCAGAGCAGCAAACATGTTCAAAATGGCC
               |||  | ||||| ||   |||   ||||    | || |  |||||||| || ||| ||
DMU09374     AACAAGGTGGAGGACGCCATCGAGTGCTACCAGCGGGCGGGCAACATGTTTAAGATGTCC
                    250       260       270       280       290       300

                  240       250       260       270       280       290
u39412.gb_pr AAAAACTGGAGTGCTGCTGGAAACGCGTTCTGCCAGGCTGCACAGCTGCACCTGCAGCTC
             ||||||||||     |||||  |    |||||| |||| |||   || |||  || |
DMU09374     AAAAACTGGACAAAGGCTGGGGAGTGCTTCTGCGAGGCGGCAACTCTACACGCGCGGGCT
                    310       320       330       340       350       360
```

FASTA on the Web

Many websites offer **FASTA** searches

- Various databases and various other services
- Be sure to use **FASTA 3**

- Each server has its limits
- Be aware that you are depending on the kindness of strangers.

Institut de Génétique Humaine, Montpellier France, GeneStream server

http://www2.igh.cnrs.fr/bin/fasta-guess.cgi

Oak Ridge National Laboratory GenQuest server

http://avalon.epm.ornl.gov/

European Bioinformatics Institute, Cambridge, UK

http://www.ebi.ac.uk/htbin/fasta.py?request

EMBL, Heidelberg, Germany

http://www.embl-heidelberg.de/cgi/fasta-wrapper-free

Munich Information Center for Protein Sequences (MIPS)
at Max-Planck-Institut, Germany

http://speedy.mips.biochem.mpg.de/mips/programs/fasta.html

Institute of Biology and Chemistry of Proteins Lyon, France

http://www.ibcp.fr/serv_main.html

Institute Pasteur, France

http://central.pasteur.fr/seqanal/interfaces/fasta.html

GenQuest at The Johns Hopkins University

http://www.bis.med.jhmi.edu/Dan/gq/gq.form.html

National Cancer Center of Japan

http://bioinfo.ncc.go.jp

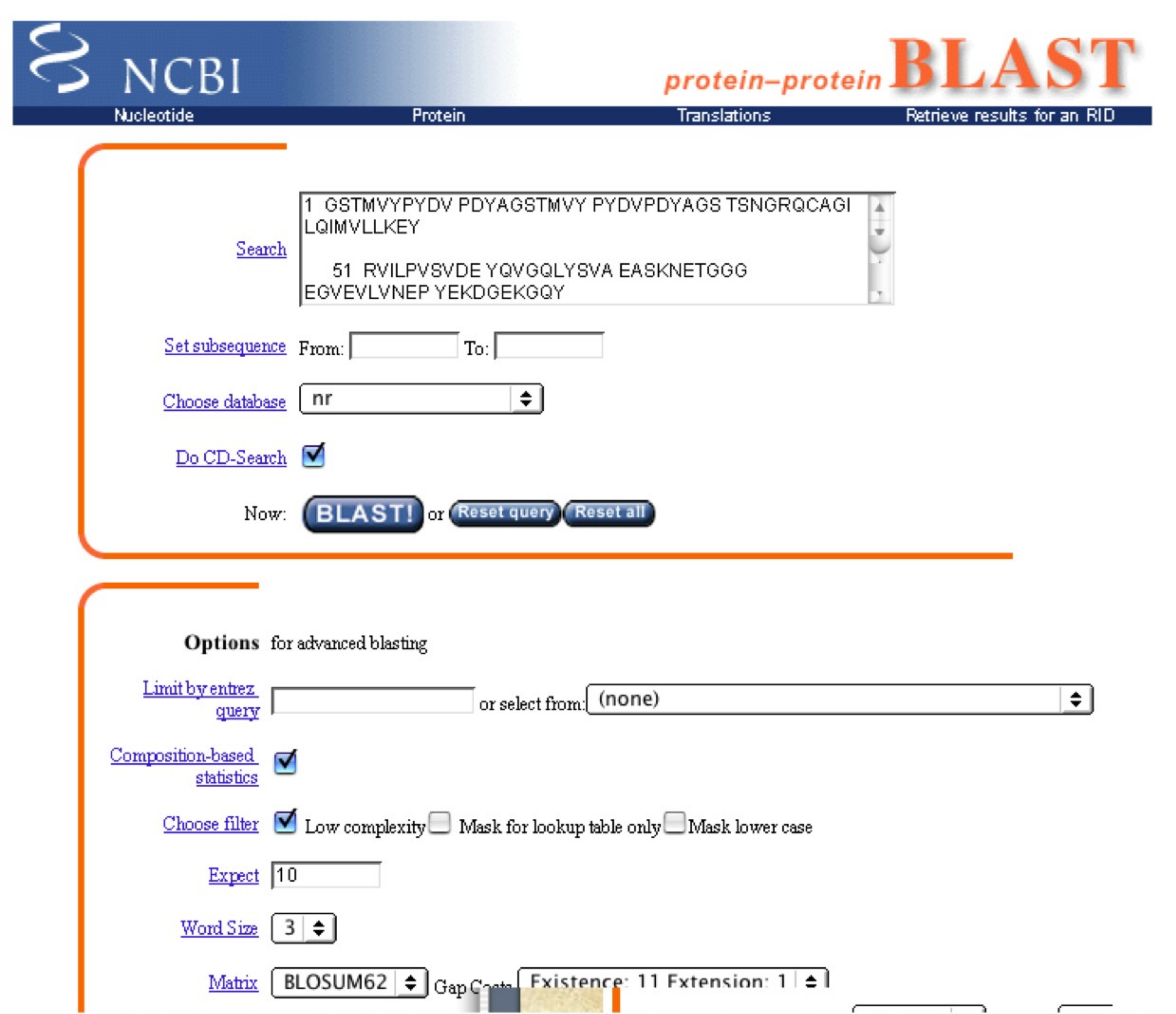
BLAST Searches GenBank

[**BLAST**= **B**asic **L**ocal **A**lignment **S**earch **T**ool]

The NCBI **BLAST** web server lets you compare your query sequence to various sections of GenBank:

- **nr** = non-redundant (main sections)
- **month** = new sequences from the past few weeks
- **ESTs**
- human, drososphila, yeast, or E.coli genomes
- proteins (by automatic translation)

- This is a <u>VERY</u> fast and powerful computer.



Search

```
1  GSTMVYPYDV PDYAGSTMVY PYDVPDYAGS TSNGRQCAGI
LQIMVLLKEY

    51  RVILPVSVDE YQVGQLYSVA EASKNETGGG
EGVEVLVNEP YEKDGEKGQY
```

Set subsequence From: ____ To: ____

Choose database: nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query ____ or select from: (none)

Composition-based statistics ☑

Choose filter ☑ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

Expect: 10

Word Size: 3

Matrix: BLOSUM62 Gap C Existence: 11 Extension: 1

Web **BLAST** runs on a big computer at NCBI

- Usually fast, but does get busy sometimes
- Fixed choices of databases
  - problems with genome data “clogging” the system
  - ESTs are not part of the default “NR” dataset
- Uses filtering of repeats (by default)
- Graphical summary of output
- Links to GenBank sequences

BLAST

- Uses word matching like FASTA
- Similarity matching of words (3 aa's, 11 bases)
  - does not require identical words.
- If no words are similar, then no alignment
  - won't find matches for very short sequences
- Does not handle gaps well
- "gapped BLAST" (BLAST 2) is better
- BLAST searches can be sent to the NCBI's server from the web or a custom client program on a personal computer or Mainframe.

Search with Protein, not DNA Sequences

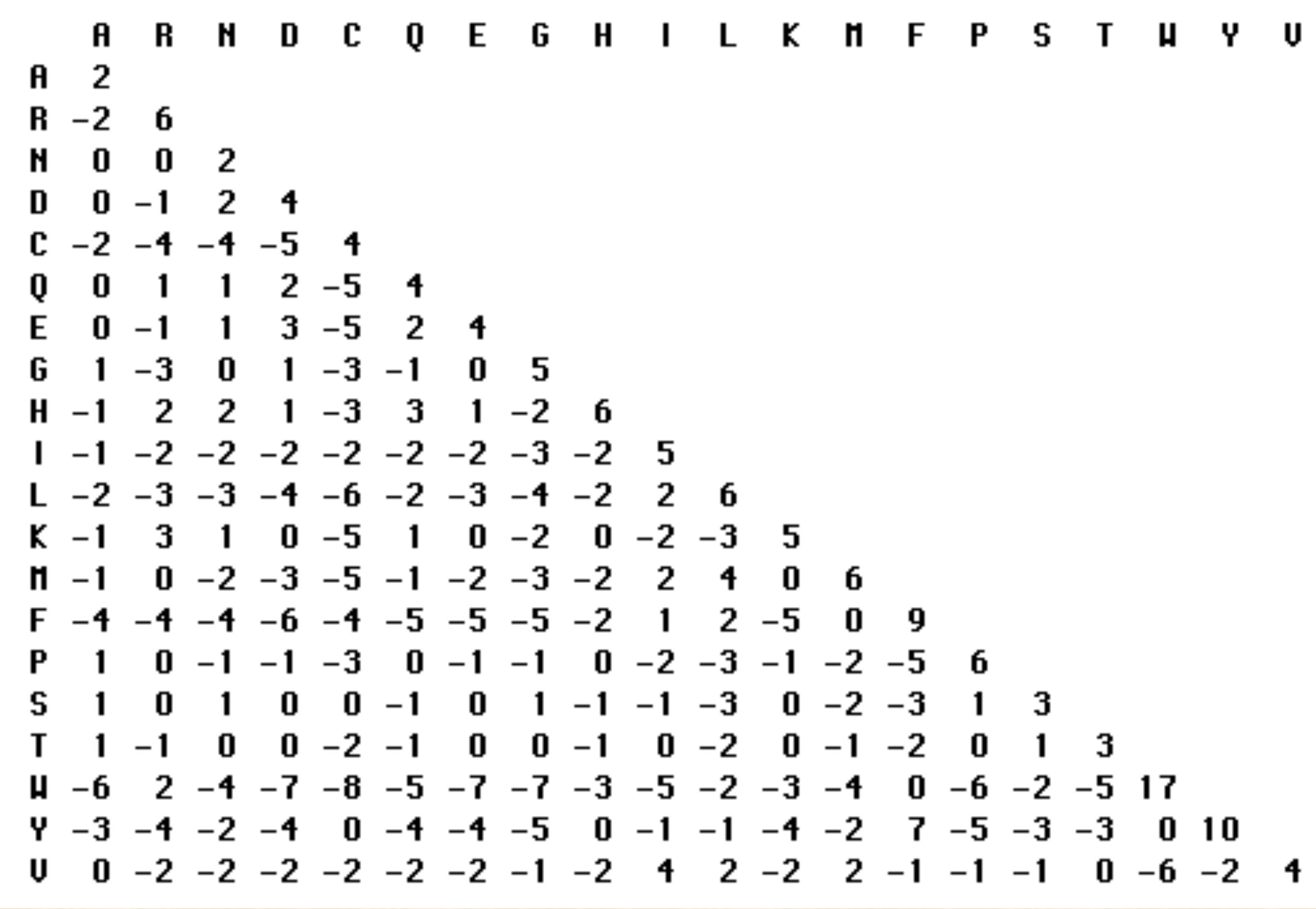
1) 4 DNA bases vs. 20 amino acids - less chance similarity

2) can have varying degrees of similarity between different AAs

- # of mutations, chemical similarity, PAM matrix

3) protein databanks are much smaller than DNA databanks

The PAM 250 scoring matrix

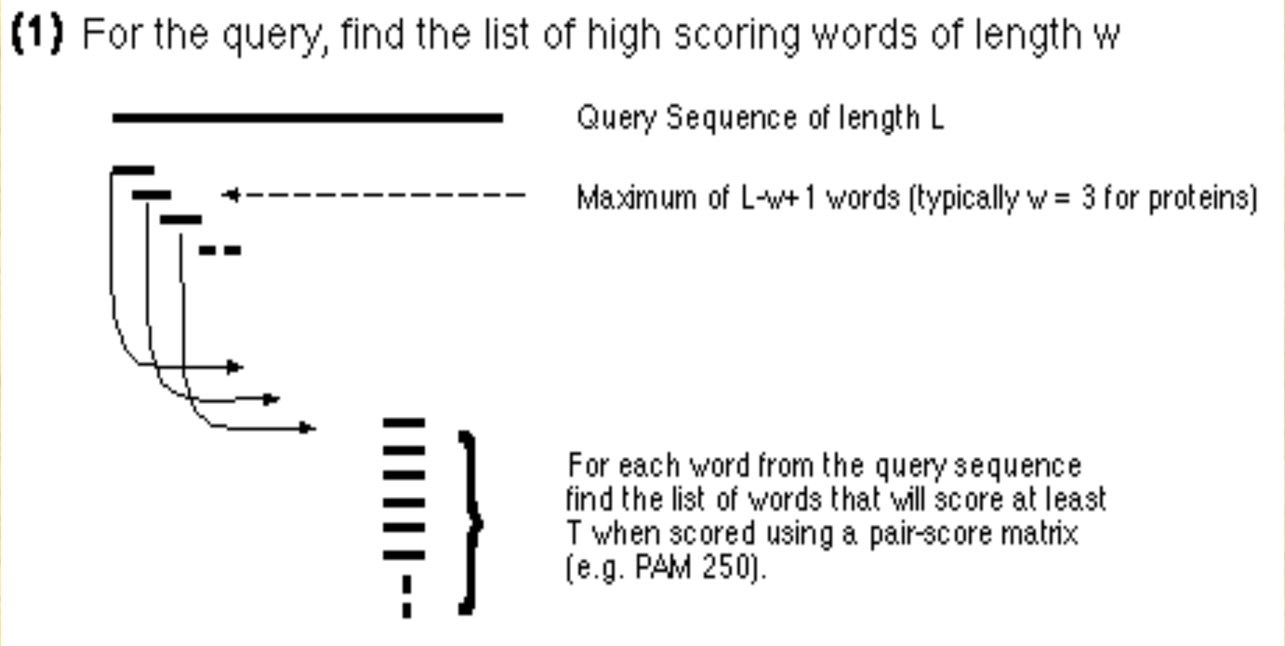


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

BLAST has Automatic Translation

- **BLASTX** makes automatic translation (in all 6 reading frames) of your DNA query sequence to compare with protein databanks
- **TBLASTN** makes automatic translation of an entire DNA database to compare with your protein query sequence
- Only make a DNA-DNA search if you are working with a sequence that does not code for protein.

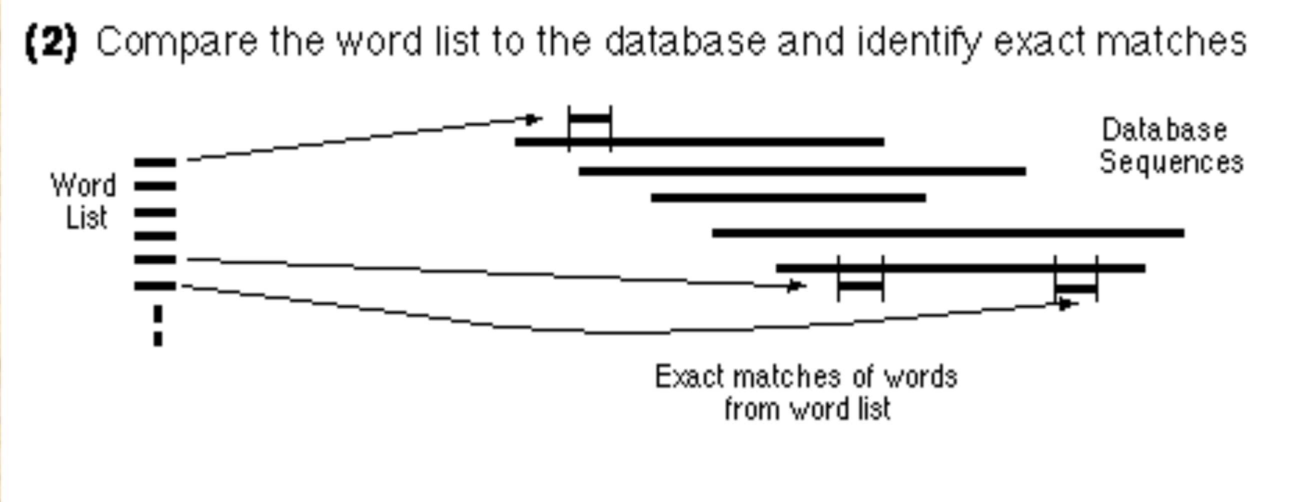
BLAST Algorithm



(1) For the query, find the list of high scoring words of length w

Query Sequence of length L
Maximum of L-w+1 words (typically w = 3 for proteins)

For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).



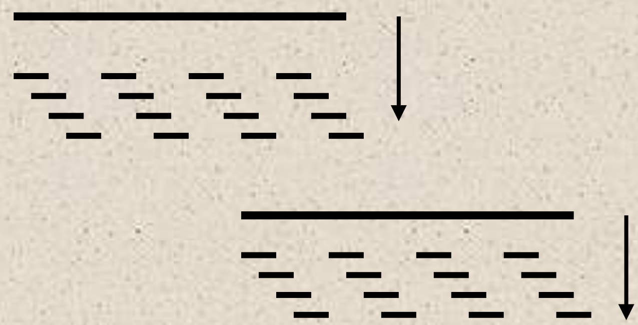
(2) Compare the word list to the database and identify exact matches

BLAST Word Matching

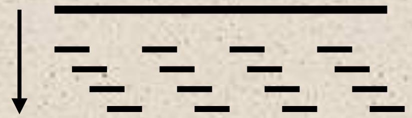
```
MEAAVKEEISVEDEAVDKNI

MEA
 EAA
  AAV
   AVK
    VKE
     KEE
      EEI
       EIS
        ISV
         ...
```

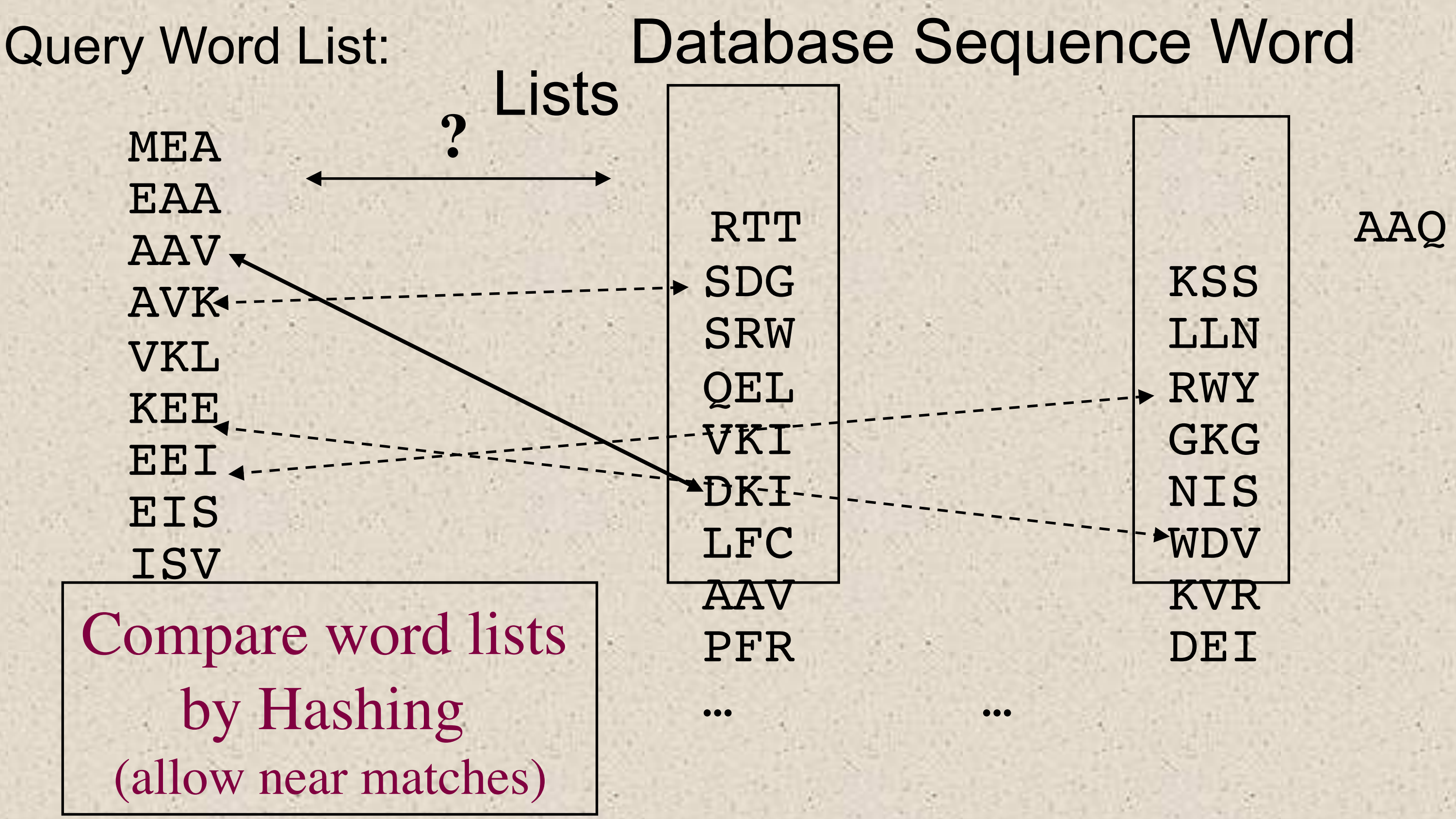
Break query into words:



Break database sequences into words:



Compare Word Lists



Query Word List:

MEA
EAA
AAV
AVK
VKL
KEE
EEI
EIS
ISV

?

Database Sequence Word Lists

RTT
SDG
SRW
QEL
VKI
DKI
LFC
AAV
PFR
...

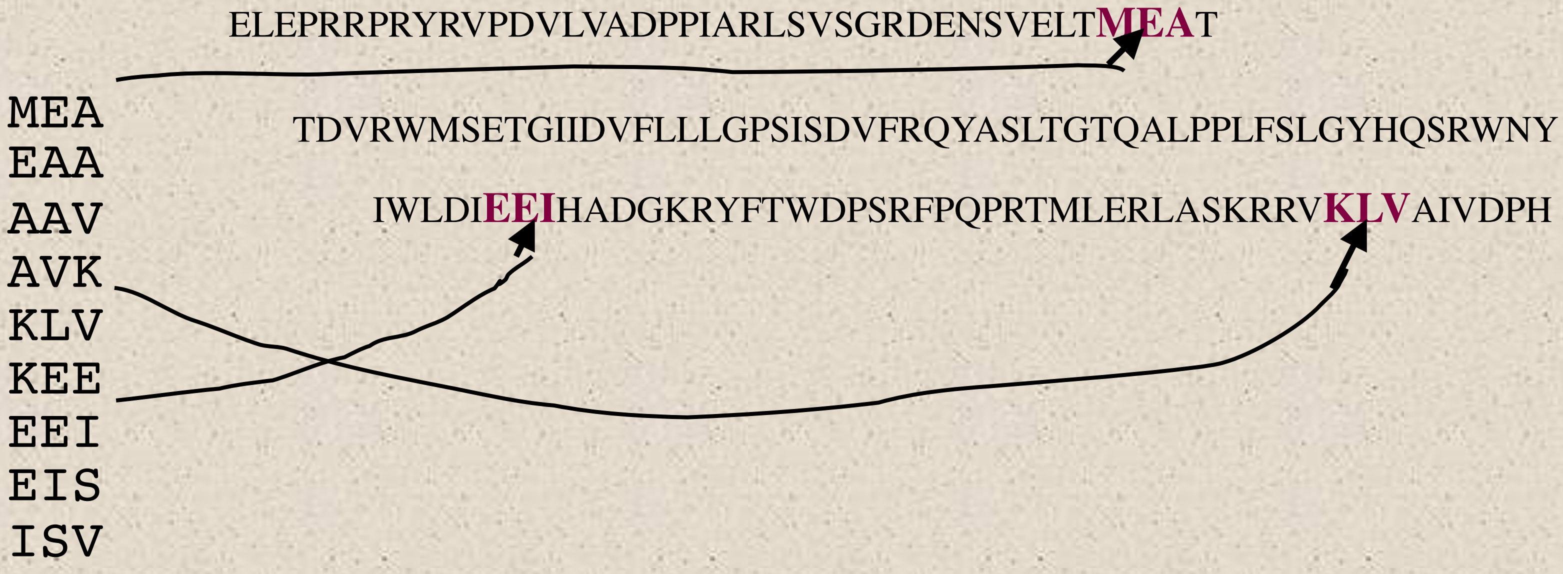
...

KSS
LLN
RWY
GKG
NIS
WDV
KVR
DEI

AAQ

Compare word lists by Hashing
(allow near matches)

Find locations of matching words in database sequences



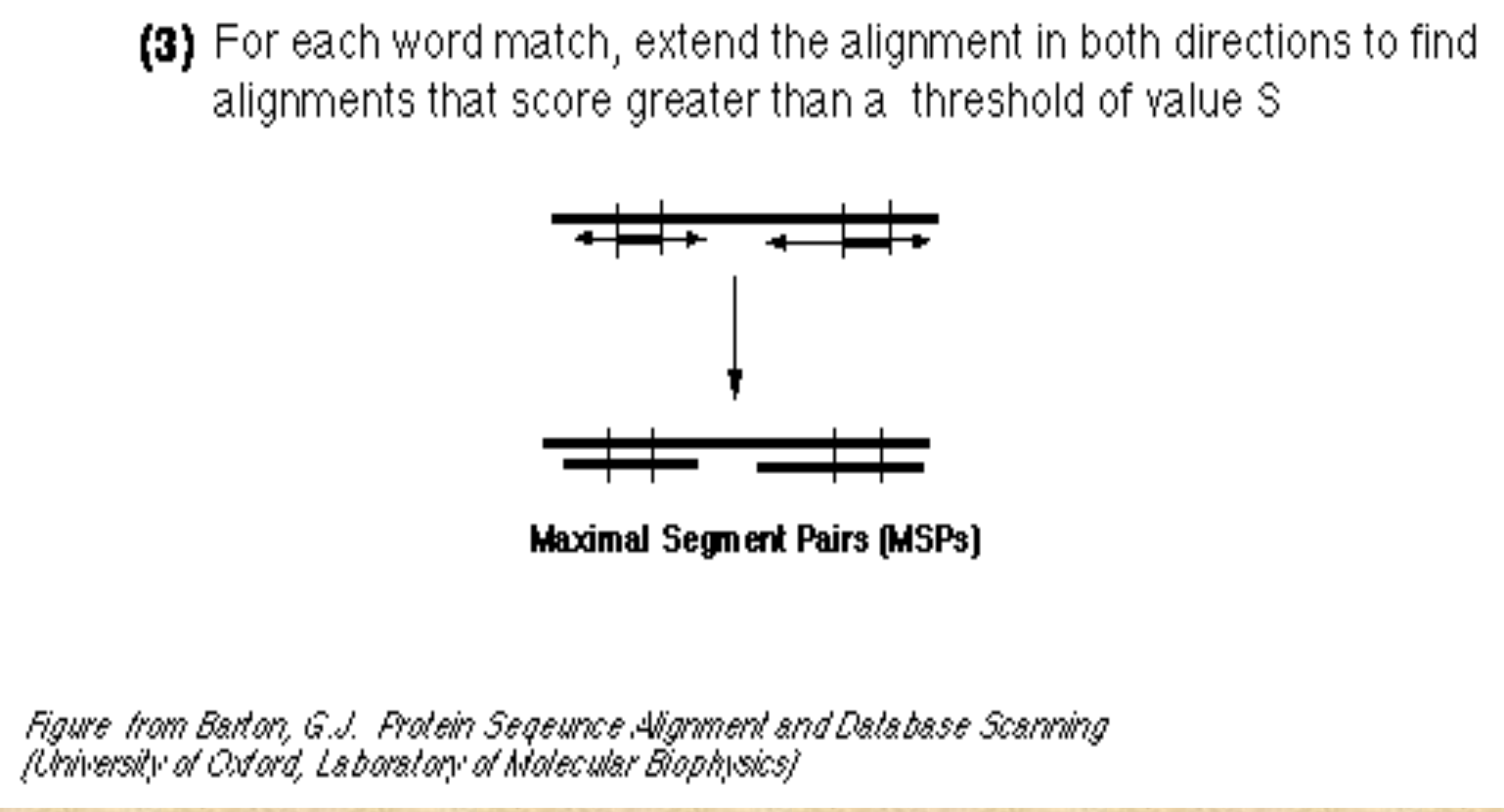
ELEPRRPRYRVPDVLVADPPIARLSVSGRDENSVELT**MEA**T

TDVRWMSETGIIDVFLLLGPSISDVFRQYASLTGTQALPPLFSLGYHQSRWNY

IWLDI**EEI**HADGKRYFTWDPSRFPQPRTMLERLASKRRV**KLV**AIVDPH

MEA
EAA
AAV
AVK
KLV
KEE
EEI
EIS
ISV

Extend hits one base at a time



(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S

Maximal Segment Pairs (MSPs)

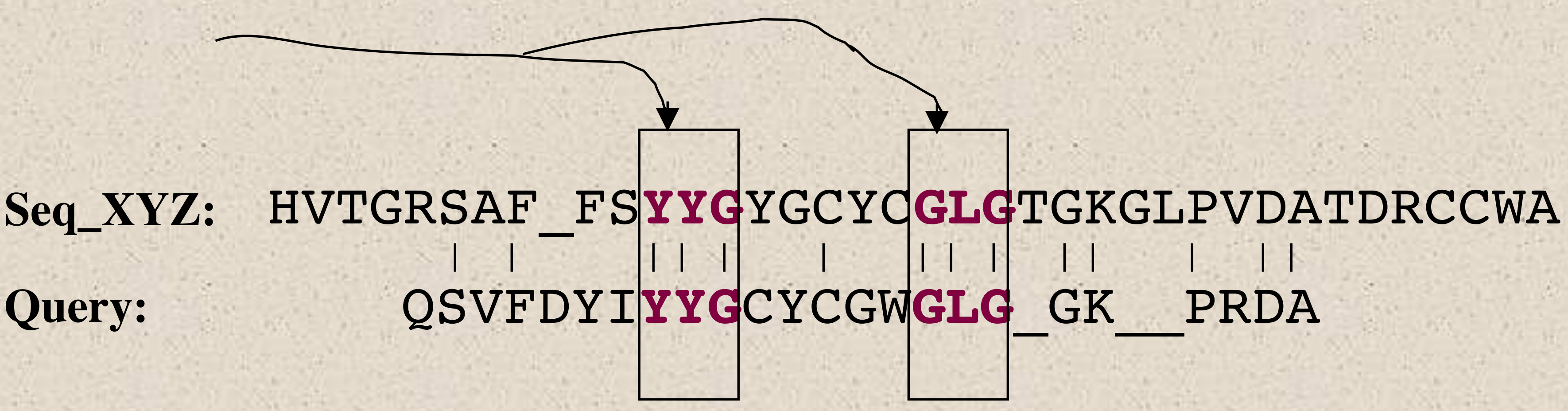
Figure from Barton, G.J. Protein Seqeunce Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)

BLAST alignments are short segments

- **BLAST** tends to break alignments into non-overlapping segments
- can be confusing
- reduces overall significance score

BLAST 2 algorithm

- The NCBI's BLAST website and **GCG** (NETBLAST) now both use BLAST 2 (also known as "gapped BLAST")
- This algorithm is more complex than the original BLAST
- It requires two word matches close to each other on a pair of sequences (i.e. with a gap) before it creates an alignment



```
Seq_XYZ:  HVTGRSAF_FSYYGYGCYCGLGTGKGLPVDATDRCCWA
               | |   |||  |  ||| ||  | ||
Query:        QSVFDYIYYGCYCGWGLG_GK__PRDA
```

E-val=10^{-13}

- Use **<u>two</u>** word matches as anchors to build an alignment between the query and a database sequence.

- Then score the alignment.

HSPs are Aligned Regions

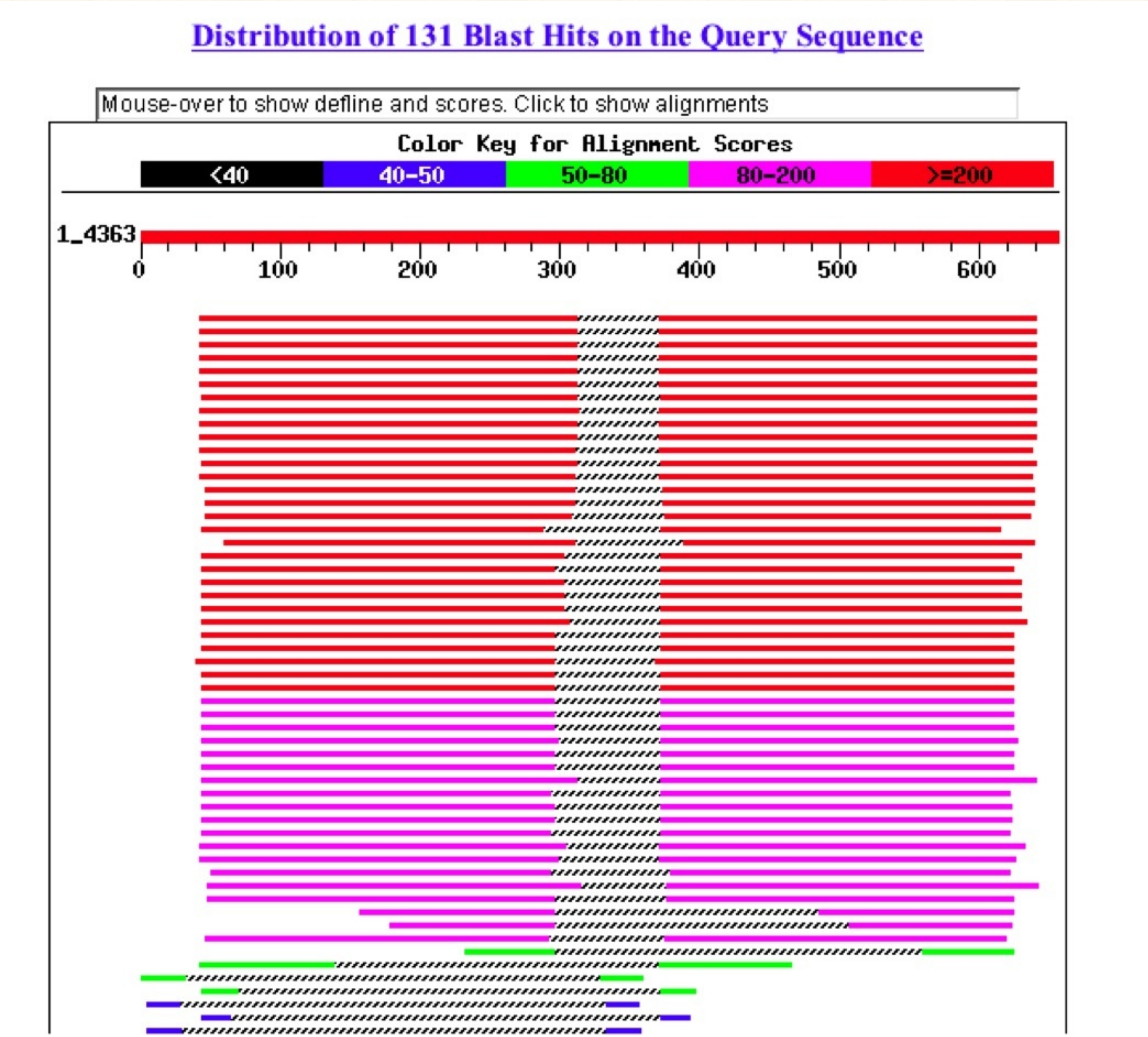
- The results of the word matching and attempts to extend the alignment are segments

  - called HSPs (High-scoring Segment Pairs)

- **BLAST** often produces several short HSPs rather than a single aligned region

```
•   >gb|BE588357.1|BE588357 194087 BARC 5BOV Bos taurus cDNA 5'.
•             Length = 369
•
•    Score =  272 bits (137), Expect = 4e-71
•    Identities = 258/297 (86%), Gaps = 1/297 (0%)
•    Strand = Plus / Plus
•
•   Query: 17  aggatccaacgtcgctccagctgctcttgacgactccacagataccccgaagccatggca 76
•              |||||||||||||||| | ||| | ||| || ||| | ||||  ||||| |||||||||
•   Sbjct: 1   aggatccaacgtcgctgcggctacccttaaccact-cgcagacccccgcagccatggcc 59
•
•   Query: 77  agcaagggcttgcaggacctgaagcaacaggtggaggggaccgcccaggaagccgtgtca 136
•              |||||||||||||||||||||||| | || ||||||||| | ||||||||||| ||| ||
•   Sbjct: 60  agcaagggcttgcaggacctgaagaagcaagtggagggggcggcccaggaagcggtgaca 119
•
•   Query: 137 gcggccggagcggcagctcagcaagtggtggaccaggccacagaggcggggcagaaagcc 196
•               |||||||| | || | ||||||||||||||| ||||||||||| || ||||||||||||
•   Sbjct: 120 tcggccggaacagcggttcagcaagtggtggatcaggccacagaagcagggcagaaagcc 179
•
•   Query: 197 atggaccagctggccaagaccacccaggaaaccatcgacaagactgctaaccaggcctct 256
•              ||||||||| | |||||||| |||||||||||||||||| ||||||||||||||||||||
•   Sbjct: 180 atggaccaggttgccaagactacccaggaaaccatcgaccagactgctaaccaggcctct 239
•
•   Query: 257 gacaccttctctgggattgggaaaaaattcggcctcctgaaatgacagcagggagac 313
•              || || ||||| ||  ||||||||||| | |||||||||||||||||| ||||||||
•   Sbjct: 240 gagactttctcgggttttgggaaaaaacttggcctcctgaaatgacagaagggagac 296
```

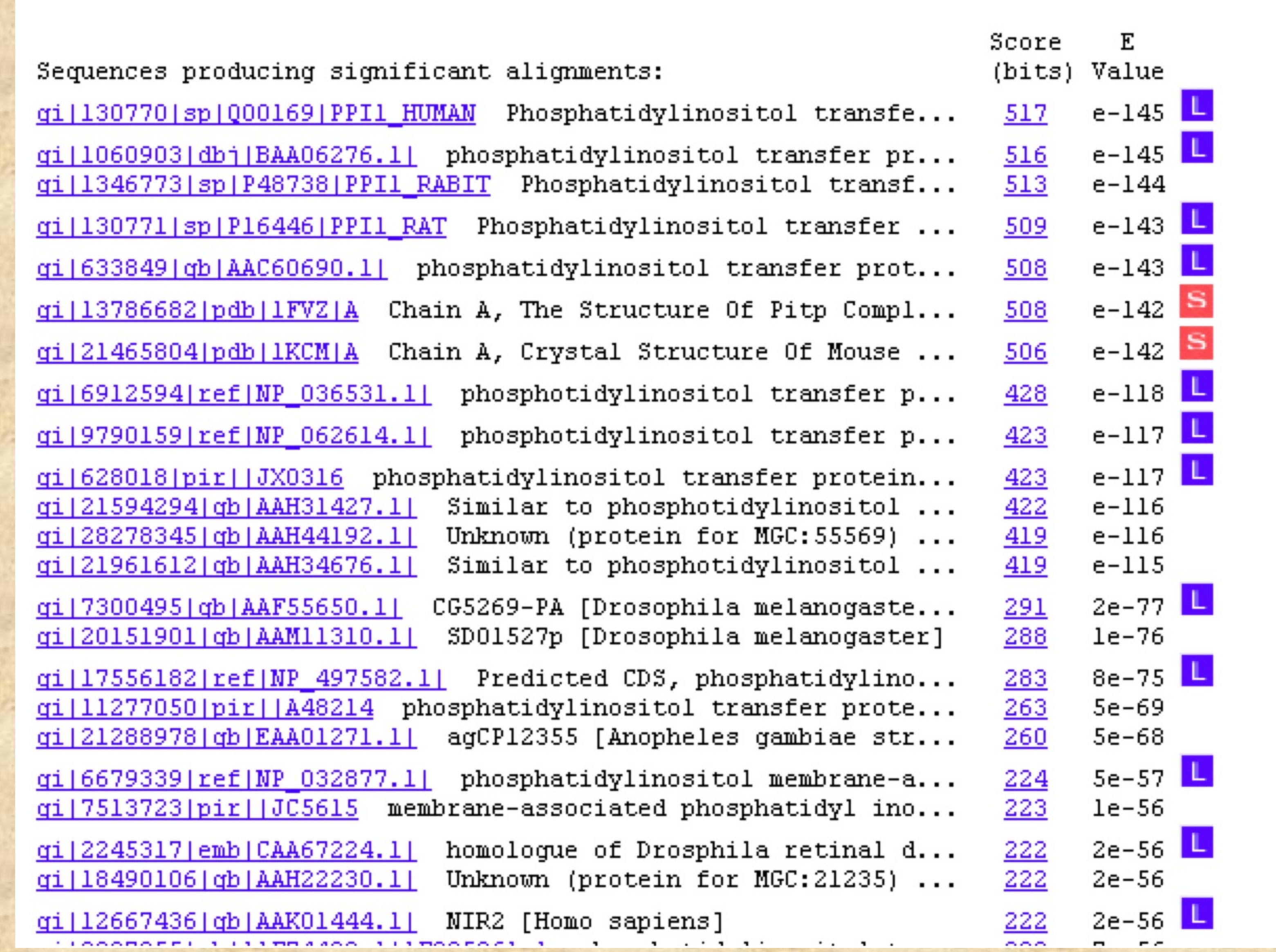
BLAST Results - Summary



Distribution of 131 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

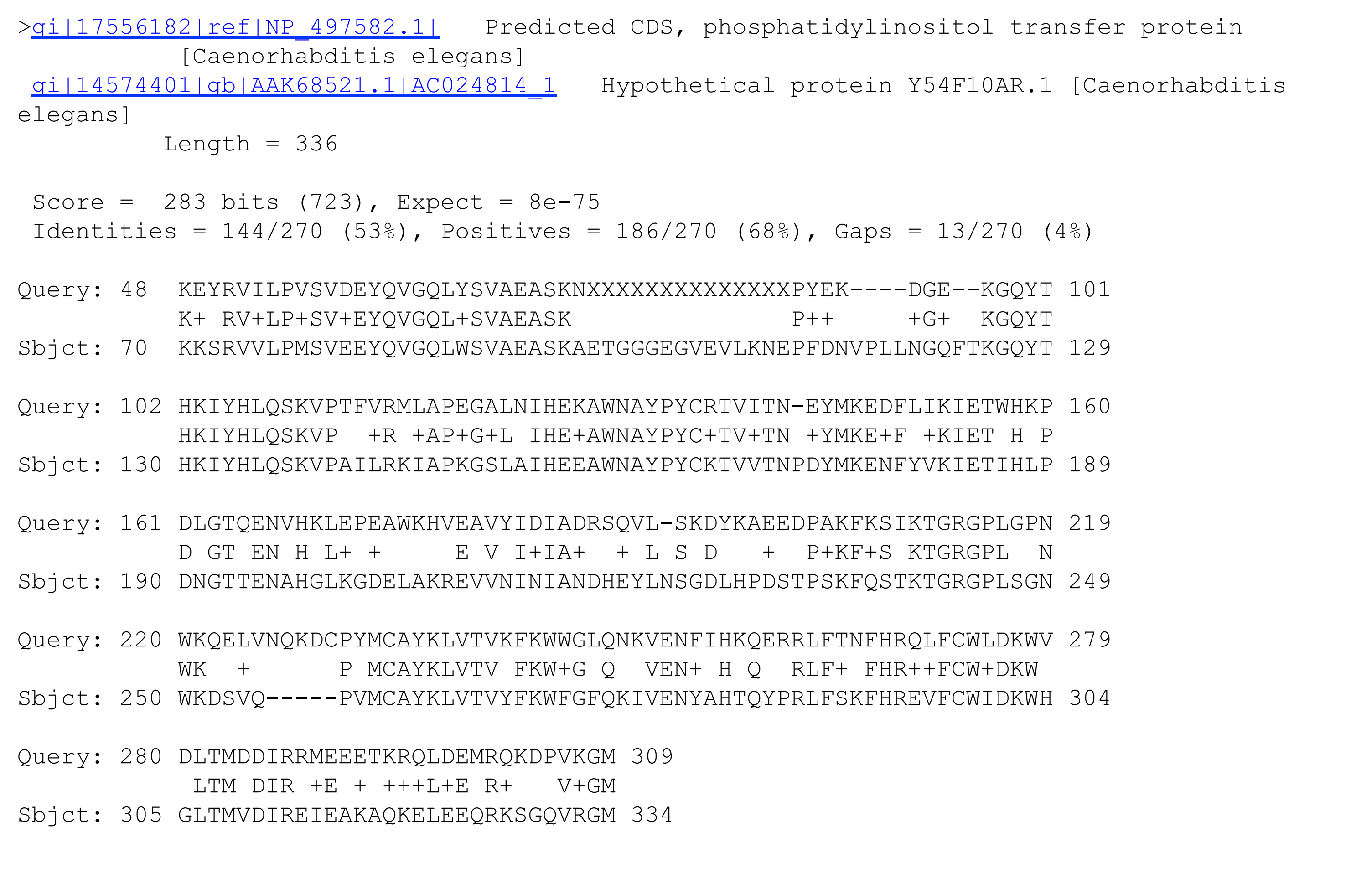
BLAST Results - List



```
                                                                     Score    E
Sequences producing significant alignments:                          (bits) Value

gi|130770|sp|Q00169|PPI1_HUMAN  Phosphatidylinositol transfe...      517   e-145  L
gi|1060903|dbj|BAA06276.1|  phosphatidylinositol transfer pr...      516   e-145  L
gi|1346773|sp|P48738|PPI1_RABIT  Phosphatidylinositol transf...      513   e-144
gi|130771|sp|P16446|PPI1_RAT  Phosphatidylinositol transfer ...      509   e-143  L
gi|633849|gb|AAC60690.1|  phosphatidylinositol transfer prot...      508   e-143  L
gi|13786682|pdb|1FVZ|A  Chain A, The Structure Of Pitp Compl...      508   e-142  S
gi|21465804|pdb|1KCM|A  Chain A, Crystal Structure Of Mouse ...      506   e-142  S
gi|6912594|ref|NP_036531.1|  phosphotidylinositol transfer p...      428   e-118  L
gi|9790159|ref|NP_062614.1|  phosphotidylinositol transfer p...      423   e-117  L
gi|628018|pir||JX0316  phosphatidylinositol transfer protein...      423   e-117  L
gi|21594294|gb|AAH31427.1|  Similar to phosphotidylinositol ...      422   e-116
gi|28278345|gb|AAH44192.1|  Unknown (protein for MGC:55569) ...      419   e-116
gi|21961612|gb|AAH34676.1|  Similar to phosphotidylinositol ...      419   e-115
gi|7300495|gb|AAF55650.1|  CG5269-PA [Drosophila melanogaste...      291   2e-77  L
gi|20151901|gb|AAM11310.1|  SD01527p [Drosophila melanogaster]       288   1e-76
gi|17556182|ref|NP_497582.1|  Predicted CDS, phosphatidylino...      283   8e-75  L
gi|11277050|pir||A48214  phosphatidylinositol transfer prote...      263   5e-69
gi|21288978|gb|EAA01271.1|  agCP12355 [Anopheles gambiae str...      260   5e-68
gi|6679339|ref|NP_032877.1|  phosphatidylinositol membrane-a...      224   5e-57  L
gi|7513723|pir||JC5615  membrane-associated phosphatidyl ino...      223   1e-56
gi|2245317|emb|CAA67224.1|  homologue of Drosphila retinal d...      222   2e-56  L
gi|18490106|gb|AAH22230.1|  Unknown (protein for MGC:21235) ...      222   2e-56
gi|12667436|gb|AAK01444.1|  NIR2 [Homo sapiens]                      222   2e-56  L
```

BLAST Results - Alignment



```
>gi|17556182|ref|NP_497582.1|   Predicted CDS, phosphatidylinositol transfer protein
           [Caenorhabditis elegans]
 gi|14574401|gb|AAK68521.1|AC024814_1   Hypothetical protein Y54F10AR.1 [Caenorhabditis
elegans]
          Length = 336

 Score =  283 bits (723), Expect = 8e-75
 Identities = 144/270 (53%), Positives = 186/270 (68%), Gaps = 13/270 (4%)

Query: 48  KEYRVILPVSVDEYQVGQLYSVAEASKNXXXXXXXXXXXXXXXPYEK----DGE--KGQYT 101
           K+ RV+LP+SV+EYQVGQL+SVAEASK               P++     +G+  KGQYT
Sbjct: 70  KKSRVVLPMSVEEYQVGQLWSVAEASKAETGGGEGVEVLKNEPFDNVPLLNGQFTKGQYT 129

Query: 102 HKIYHLQSKVPTFVRMLAPEGALNIHEKAWNAYPYCRTVITN-EYMKEDFLIKIETWHKP 160
           HKIYHLQSKVP  +R +AP+G+L IHE+AWNAYPYC+TV+TN +YMKE+F +KIET H P
Sbjct: 130 HKIYHLQSKVPAILRKIAPKGSLAIHEEAWNAYPYCKTVVTNPDYMKENFYVKIETIHLP 189

Query: 161 DLGTQENVHKLEPEAWKHVEAVYIDIADRSQVL-SKDYKAEEDPAKFKSIKTGRGPLGPN 219
           D GT EN H L+ +     E V I+IA+  + L S D   +  P+KF+S KTGRGPL  N
Sbjct: 190 DNGTTENAHGLKGDELAKREVVNINIANDHEYLNSGDLHPDSTPSKFQSTKTGRGPLSGN 249

Query: 220 WKQELVNQKDCPYMCAYKLVTVKFKWWGLQNKVENFIHKQERRLFTNFHRQLFCWLDKWV 279
           WK  +      P MCAYKLVTV FKW+G Q  VEN+ H Q  RLF+ FHR++FCW+DKW
Sbjct: 250 WKDSVQ-----PVMCAYKLVTVYFKWFGFQKIVENYAHTQYPRLFSKFHREVFCWIDKWH 304

Query: 280 DLTMDDIRRMEEETKRQLDEMRQKDPVKGM 309
            LTM DIR +E + +++L+E R+   V+GM
Sbjct: 305 GLTMVDIREIEAKAQKELEEQRKSGQVRGM 334
```
]

FASTA/BLAST **Statistics**

- E() value is equivalent to standard P value
- Significant if E() < 0.05 (smaller numbers are more significant)
  - The E-value represents the likelihood that the observed alignment is due to chance alone. A value of 1 indicates that an alignment this good would happen by chance with any random sequence searched against this database.

BLAST is Approximate

- BLAST makes similarity searches very quickly because it takes shortcuts.
  - looks for short, nearly identical “words” (11 bases)
- It also makes errors
  - misses some important similarities
  - makes many incorrect matches
    - easily fooled by repeats or skewed composition

Interpretation of output

- very low E() values (< e-100) are homologs or identical genes
- moderate E() values (~ e-50) are related genes
- long list of gradually declining of E() values indicates a large gene family
- long regions of moderate similarity are more significant than short regions of high identity

Biological Relevance

- It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the mismatches conservative ones?
- Are the matching regions important structural components of the genes or just introns and flanking regions?

Borderline similarity

- What to do with matches with E() values in the 0.5 -1.0 range?
- this is the **“Twilight Zone”**
- retest these sequences and look for related hits (not just your original query sequence)
- similarity is transitive:

if **A~B** and **B~C**, then **A~C**

Advanced Similarity Techniques

Automated ways of using the results of one search to initiate multiple searches

- **INCA** (**I**terative **N**eighborhood **C**luster **A**nalysis) **http://itsa.ucsf.edu/~gram/home/inca/**
  - Takes results of one BLAST search, does new searches with each one, then combines all results into a single list
  - JAVA applet, compatibility problems on some computers
- **PSI BLAST** **http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html**
  - Creates a "position specific scoring matrix" from the results of one BLAST search
  - Uses this matrix to do another search
  - builds a family of related sequences
  - can't trust the resulting e-values

PSI BLAST

- Starts with a single BLAST search
    - only works on PROTEIN
- Finds matches: builds a new scoring matrix just for this set of sequences
- Use the new matrix to search for more distant matches
- Repeat
- Results are only as good as your intial set of sequences used to build the matrix

Database to Search

- The biggest factor that affects the results of a similarity search, is …obviously… what database you search
- Choose to search PROTEIN databases whenever possible
  - Smaller = less redundant = higher e-values
  - Non-identical letters have information (scoring matrix)

Comprehensive vs Annotated

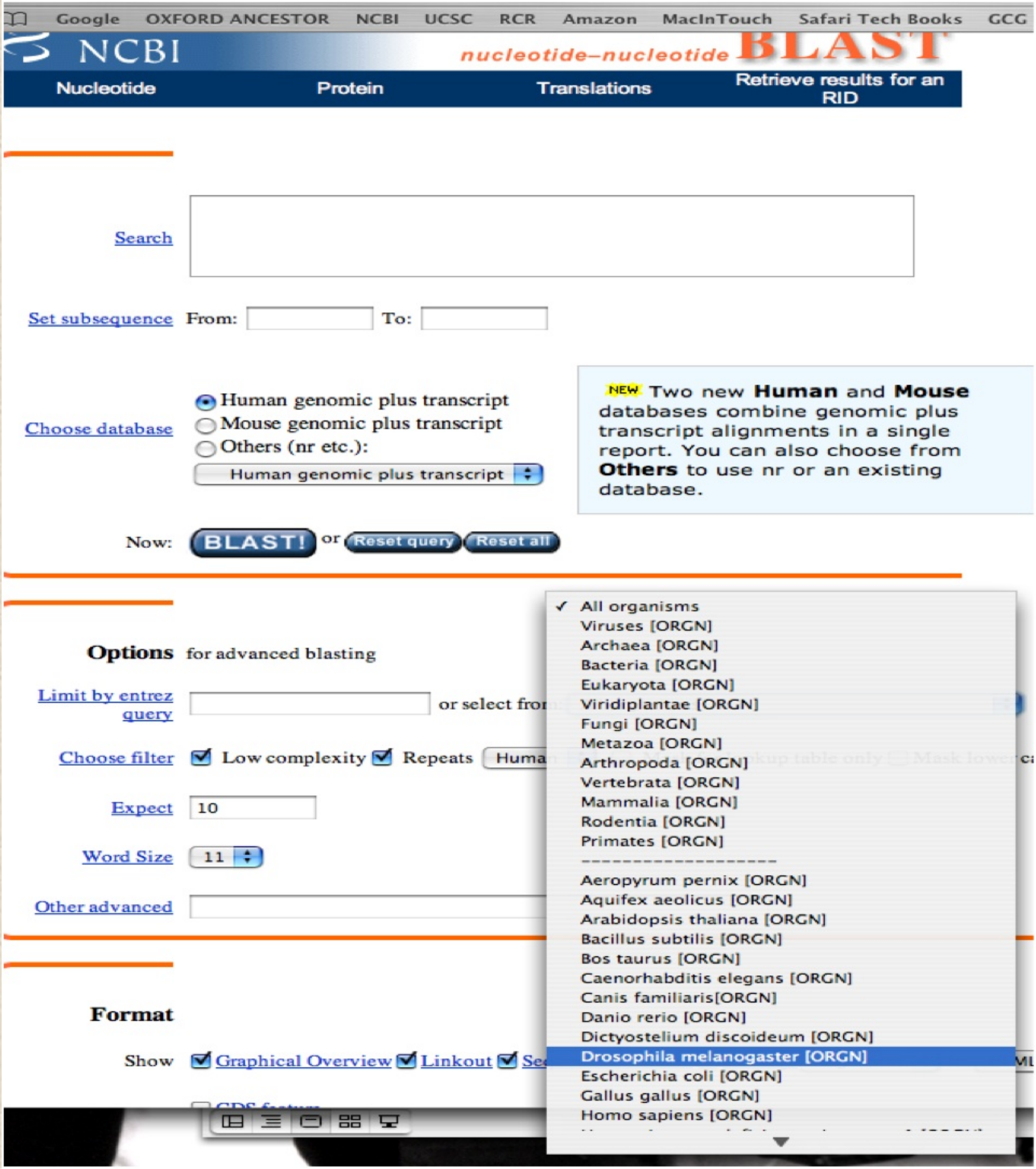
- It is NOT always best to search the biggest, most comprehensive database
- What have you learned when your cloned sequence matches a "hypothetical gene?"
- RefSeq is the best annotated DNA database
- SwissProt is the best annotated protein database

What are you looking for?

- Usually you want to search annotated genes
- If you don't find anything, you might want to search ESTs (sequences of mRNA fragments)
- ESTs are not included in the default "nr" GenBank database

Limit by species

- If you know your sequence is from one species
- Or you want to limit your search to just that species…
- use the ENTREZ limits feature



Set subsequence From: To:

Choose database

● Human genomic plus transcript

○ Mouse genomic plus transcript

○ Others (nr etc.):

Human genomic plus transcript

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: ✓ All organisms

Choose filter ☑ Low complexity ☑ Repeats Huma

Expect 10

Word Size 11

Other advanced

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Se

□ CDS feature

Filters

- BLAST is easily fooled by repeats and low complexity sequence (enriched in a few letters = DNA microsatellites, common acidic, basic or proline-rich regions in proteins)
- Default filters remove low complexity from protein searches and known repeats (ie. *Alu*) from DNA searches
- Removes the problem sequences before running the BLAST search
- You can turn off the filters to get true alignments and e-values (`"lookup only"`)

Size Matters

- Short sequences can't get good e-values
- What is the probability of finding a 12 base fragment in a "random" genome?

  4^{12} = 16,777,216 (once per 16 million bases)

- What length DNA fragment is needed to define a unique location in the genome?

  4^{16} = 4,294,967,296 (4 billion bases)

- So, what is the best e-value you can get for a 16 base fragment?

Word size

- BLAST uses a default word size of 11 bases for DNA
- Short sequences will have few words
- Low quality sequence might have a sequencing error in every word
- "MegaBlast" uses very large words (28)
  - – allows for fast mRNA > genome alignment
  - – allows huge sequences to be use as query
- "Search for short, nearly exact matches"
    - word size = 7, expect = 1000

Batch BLAST

- What if you need to do a LOT of BLAST searches?

- NCBI www BLAST server will accept a FASTA file with multiple sequences
- NCBI has a BLAST client program:

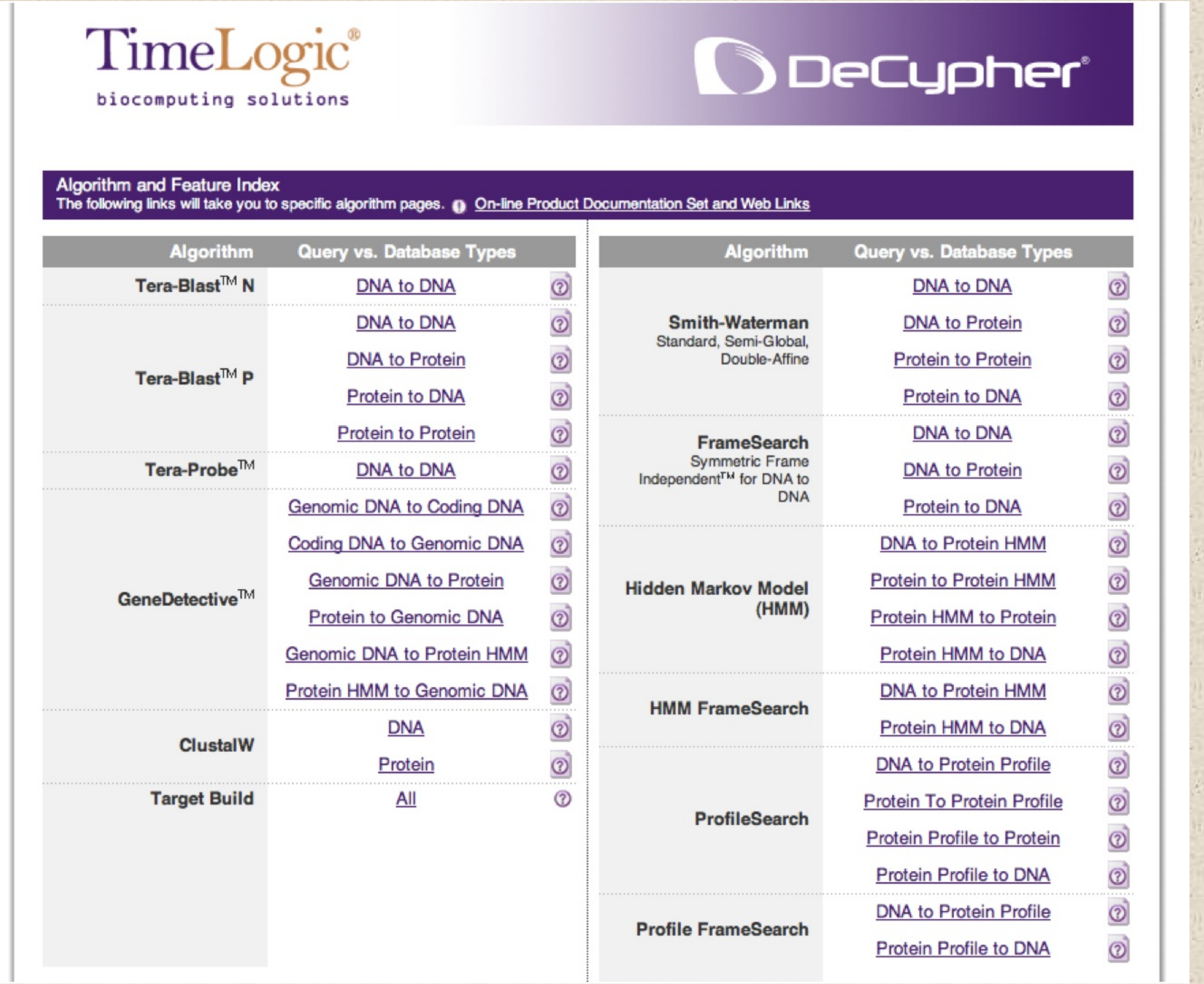
  blastcl3 (Unix, Windows, and Mac)
- NETBLAST is a scriptable BLAST client in GCG package

Accelerated BLAST

- The BLAST algorithm can run on special parallel computing hardware
- At NYU, the RCR runs a super BLAST server:

  http://codequest.med.nyu.edu

  Can create custom databases for your project



TimeLogic®
biocomputing solutions

DeCypher®

Algorithm and Feature Index

The following links will take you to specific algorithm pages. ⓘ On-line Product Documentation Set and Web Links

Algorithm	Query vs. Database Types
Tera-Blast™ N	DNA to DNA
Tera-Blast™ P	DNA to DNA
	DNA to Protein
	Protein to DNA
	Protein to Protein
Tera-Probe™	DNA to DNA
GeneDetective™	Genomic DNA to Coding DNA
	Coding DNA to Genomic DNA
	Genomic DNA to Protein
	Protein to Genomic DNA
	Genomic DNA to Protein HMM
	Protein HMM to Genomic DNA
ClustalW	DNA
	Protein
Target Build	All

Algorithm	Query vs. Database Types
Smith-Waterman Standard, Semi-Global, Double-Affine	DNA to DNA
	DNA to Protein
	Protein to Protein
	Protein to DNA
FrameSearch Symmetric Frame Independent™ for DNA to DNA	DNA to DNA
	DNA to Protein
	Protein to DNA
Hidden Markov Model (HMM)	DNA to Protein HMM
	Protein to Protein HMM
	Protein HMM to Protein
	Protein HMM to DNA
HMM FrameSearch	DNA to Protein HMM
	Protein HMM to DNA
ProfileSearch	DNA to Protein Profile
	Protein To Protein Profile
	Protein Profile to Protein
	Protein Profile to DNA
Profile FrameSearch	DNA to Protein Profile
	Protein Profile to DNA

Lots of Results

- Batch or acclerated BLAST searches produce lots of results files.
- What to do with them?
- BlastReport2 is a Perl script from NCBI to sort out results from a batch BLAST.

"BlastReport2 is a perl script that reads the output of Blastcl3, reformats it for ease of use and eliminates useless information."

BLAST Parser

- Hundreds of different people have written programs to sort BLAST results

  (including myself)

- Better to use a common code base
- BioPerl is a collection of public Perl modules including several BLAST parsers

ESTs have frameshifts

- How to search them as proteins?
- Can use TBLASTN but this breaks each frame-shifted region into its own little protein
- GCG FRAMESEARCH is killer slow

  (uses an extended version of the Smith-Waterman algorithm)
- FASTX (DNA vs. protein database) and TFASTX (protein vs. DNA database) search for similarity taking account of frameshifts

Genome Alignment

- How to match a protein or mRNA to genomic sequence?
  - There is a Genome BLAST server at NCBI
  - Each of the Genome websites has a similar search function
- What about introns?
  - An intron is penalized as a gap, or each exon is treated as a separate alignment with its own e-score
  - Need a search algorithm that looks for consensus intron splice sites and points in the alignment where similarity drops off.

Sim4 is for mRNA -> DNA Alignment

- *Florea L, Hartzell G, Zhang Z, Rubin GM, **Miller W**. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **1998** 8:967-74*
- This is a fairly new program (1998) as compared to **BLAST** and **FASTA**
- It is written for **UNIX** (of course), but there is a web server (and it is used in many other 'genome analysis' tools): **http://pbil.univ-lyon1.fr/sim4.html**
- Finds best set of segments of local alignment with a preference for fragments that end with splice-site recognition signals **(GT-AG, CT-AC)**

More Genome Alignment

- **Est2Genome**: like it says, compares an EST to genome sequence)

http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html

- **GeneWise**: Compares a protein (or motif) to genome sequence

http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml

What program to use for searching?

1) **BLAST** is fastest and easily accessed on the Web

- limited sets of databases
- nice translation tools (BLASTX, TBLASTN)

2) **FASTA**

precise choice of databases

- more sensitive for DNA-DNA comparisons
- **FASTX** and **TFASTX** can find similarities in sequences with frameshifts

3) Smith-Waterman - slower, but more sensitive

- known as a “rigorous” or “exhaustive” search
- **SSEARCH** in **GCG** and standalone **FASTA**

Smith-Waterman searches

- A more sensitive brute force approach to searching
- **much** slower than **BLAST** or **FASTA**
- uses dynamic programming
- **SSEARCH** is a **GCG** program for Smith-Waterman searches
- **WATER** is an **EMBOSS** program for Smith-Waterman searches

Smith-Waterman on the Web

- The **EMBL** offers a service know as **BLITZ**, which actually runs an algorithm called **MPsrch** on a dedicated **MassPar** massively parallel super-computer.

  http://www.ebi.ac.uk/bic_sw/

- The Weizmann Institute of Science offers a service called the BIOCCELERATOR provided by **Compugen** Inc.

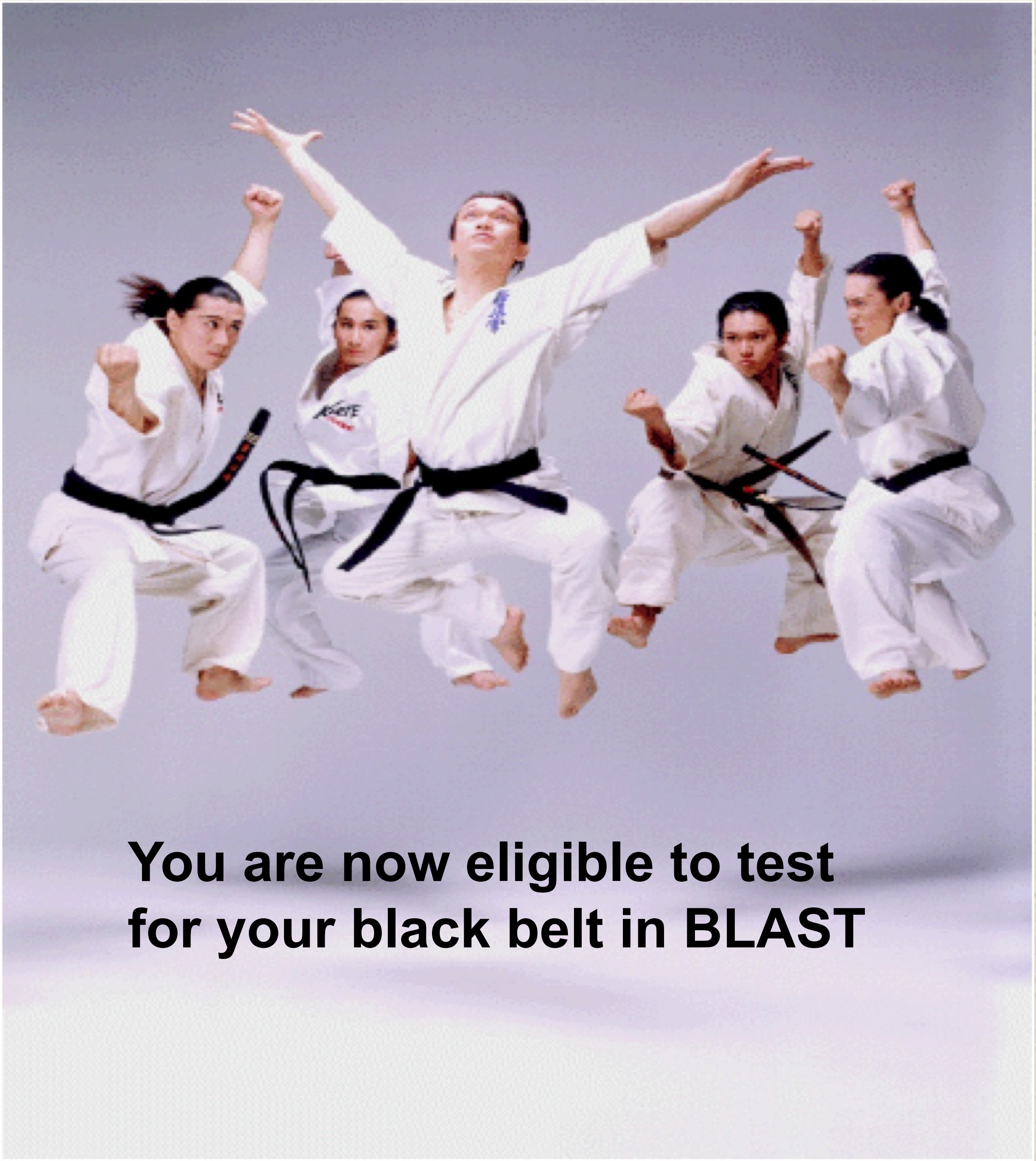
http://sgbcd.weizmann.ac.il:80/cgi-bin/genweb/main.cgi

Strategies for similarity searching

1) Web, PC program, **GCG**, or custom client?

2) Start with smaller, better annotated databases (limit by taxonomic group if possible)

3) Search **protein** databases (use translation for DNA seqs.) unless you have non-coding DNA



You are now eligible to test for your black belt in BLAST