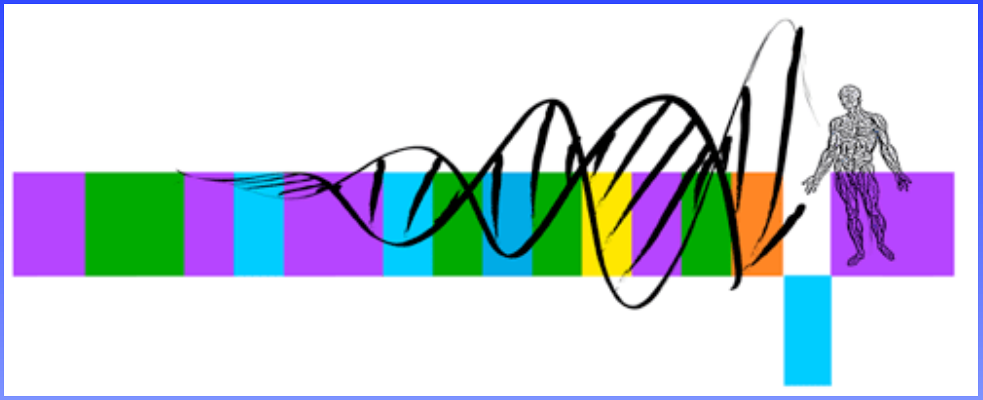


Bioinformatics for Biologists

Sequence Analysis: Part I. Pairwise alignment and database searching

Fran Lewitter, Ph.D.
Director
Bioinformatics & Research Computing
Whitehead Institute

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods

Topics to Cover

- Introduction
  - Why do alignments?
  - A bit of history
  - Definitions
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods

Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. *Science* 221:275-277, 1983.

Evolutionary Basis of Sequence Alignment

- ***Similarity*** - observable quantity, such as percent identity
- ***Homology*** - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this

Some Definitions

- An ***alignment*** is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.
- An ***optimal alignment*** is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.

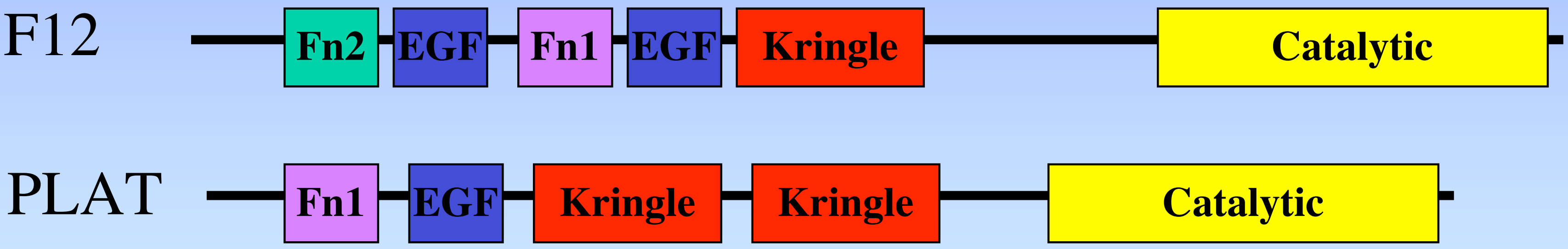
Alignment Methods

- ***Global alignment*** - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.
- ***Local alignment*** - Smith-Waterman (1981) is a modification of the dynamic programming algorithm giving the highest scoring local match between two sequences.

Alignment Methods

Global vs Local

Modular proteins



Local vs Global Alignment

```
L G P S S K Q T G K G S - S R I W D N
L N - I T K S A G K G A I M R L G D A
```

GLOBAL

```
- - - - - - - T G K G - - - - - - - -
- - - - - - - A G K G - - - - - - - -
```

LOCAL

From Mount, *Bioinformatics, 2004, pg 71*

Possible Alignments

```
A:   T C A G A C G A G T G
B:   T C G G A G C T G

I.   T C A G A C G A G T G

     T C G G A - - G C T G

II.  T C A G A C G A G T G

     T C G G A - G C - T G

III. T C A G A C G A G T G

     T C G G A - G - C T G
```

Topics to Cover

- Introduction
- Scoring alignments
  - Nucleotide vs Proteins
  - Definitions
- Alignment methods
- Significance of alignments
- Database searching methods

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 5
- Gap extension = -2

```
  T   C   A   G   A   C   G   A   G   T   G
  T   C   G   G   A   -   -   G   C   T   G
_____________________________________________
 +1  +1  -1  +1  +1  -5  -2  -1  -1  +1  +1  =  -4
```

Possible Alignments

```
A:T C A G A C G A G T G
B:T C G G A G C T G
```

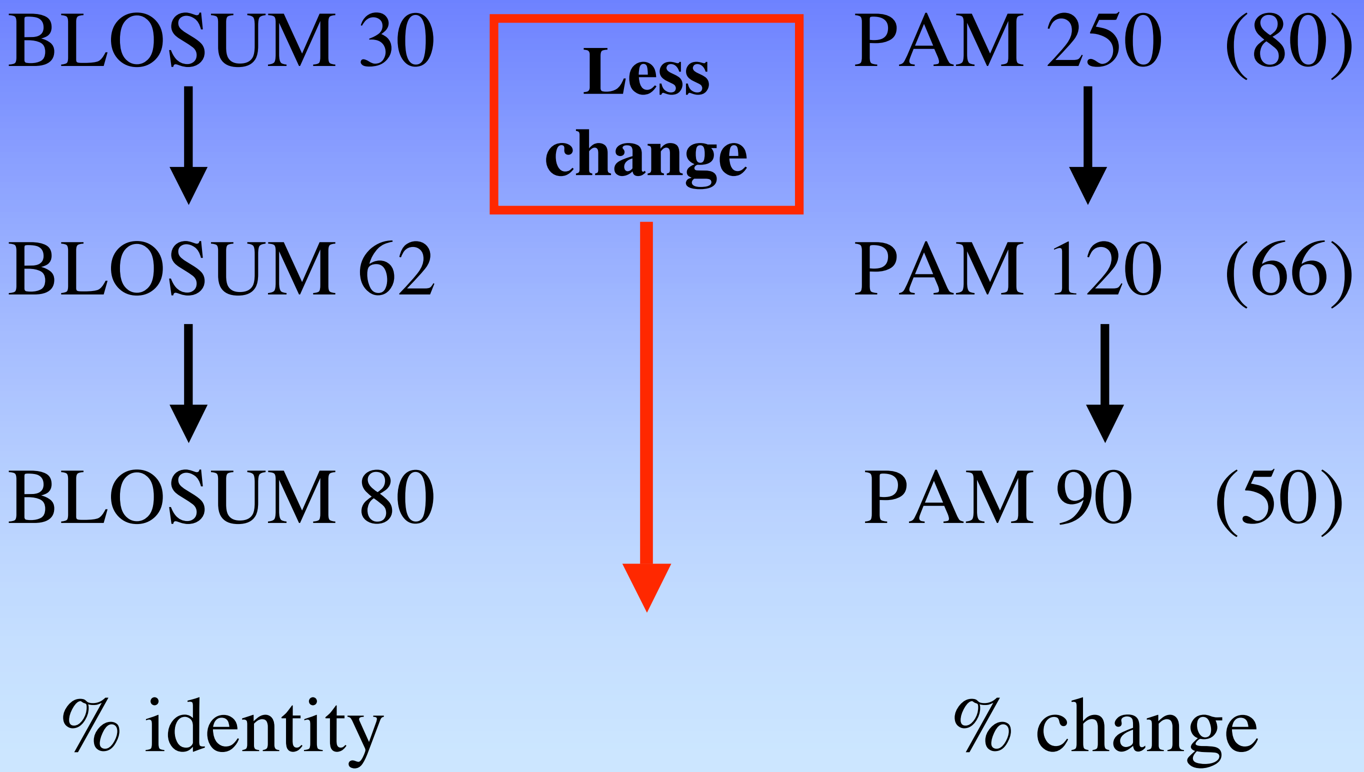
```
I.   T C A G A C G A G T G
                               -4
     T C G G A - - G C T G

II.  T C A G A C G A G T G
                               -5
     T C G G A - G C - T G

III. T C A G A C G A G T G
                               -5
     T C G G A - G - C T G
```

Amino Acid Substitution Matrices

- ***PAM*** **-** point accepted mutation based on *global* alignment [evolutionary model]

- ***BLOSUM*** - block substitutions based on *local* alignments [similarity among conserved sequences]

Substitution Matrices



BLOSUM 30

BLOSUM 62

BLOSUM 80

% identity

Less change

PAM 250 (80)

PAM 120 (66)

PAM 90 (50)

% change

Part of BLOSUM 62 Matrix

	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	

$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

Part of PAM 250 Matrix

	C	S	T	P	A	G	N
C	12						
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	

$$\textbf{Log-odds} = \frac{\textbf{pair in homologous proteins}}{\textbf{pair in unrelated proteins by chance}}$$

Gap Penalties

- ***Insertion and Deletions*** (indels)
- ***Affine gap costs*** - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length

Scoring for BLAST 2 Sequences

```
Score = 94.0 bits (230), Expect = 6e-19
Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTLSGAPCQPWASEATYRNVTAEQ---AR 256
           Y+  FC      +  + CY G G +YRG    T SGA C PW S      V   Q   A+
Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKVYTAQNPSAQ 257

Query: 257 NWGLGGHAFCRNPDNDIRPWCFVLNRDRLSWEYCDLAQCQT 297
             GLG H +CRNPD D +PWC VL   RL+WEYCD+  C T
Sbjct: 258 ALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCST 298
```

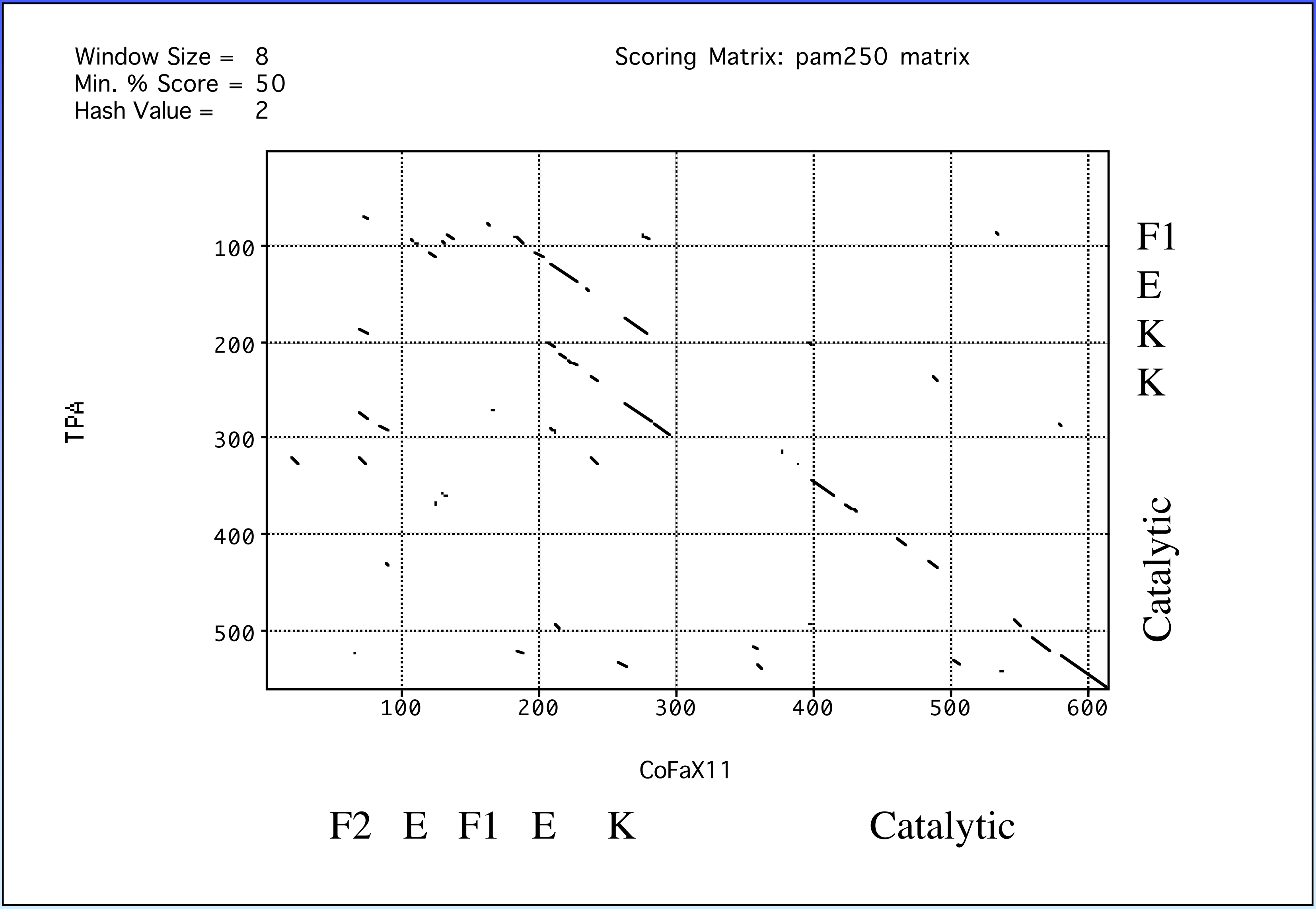
Based on BLOSUM62

```
Position  1: Y - Y =   7
Position  2: T - S =   1
Position  3: G - S =   0
Position  4: P - E =  -1
          . . .
Position  9: - - P = -11
Position 10: - - A =  -1
          . . .
                 Sum     230
```

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
  - Dot matrix analysis
  - Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  - Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST, BLAT)
- Significance of alignments
- Database searching methods

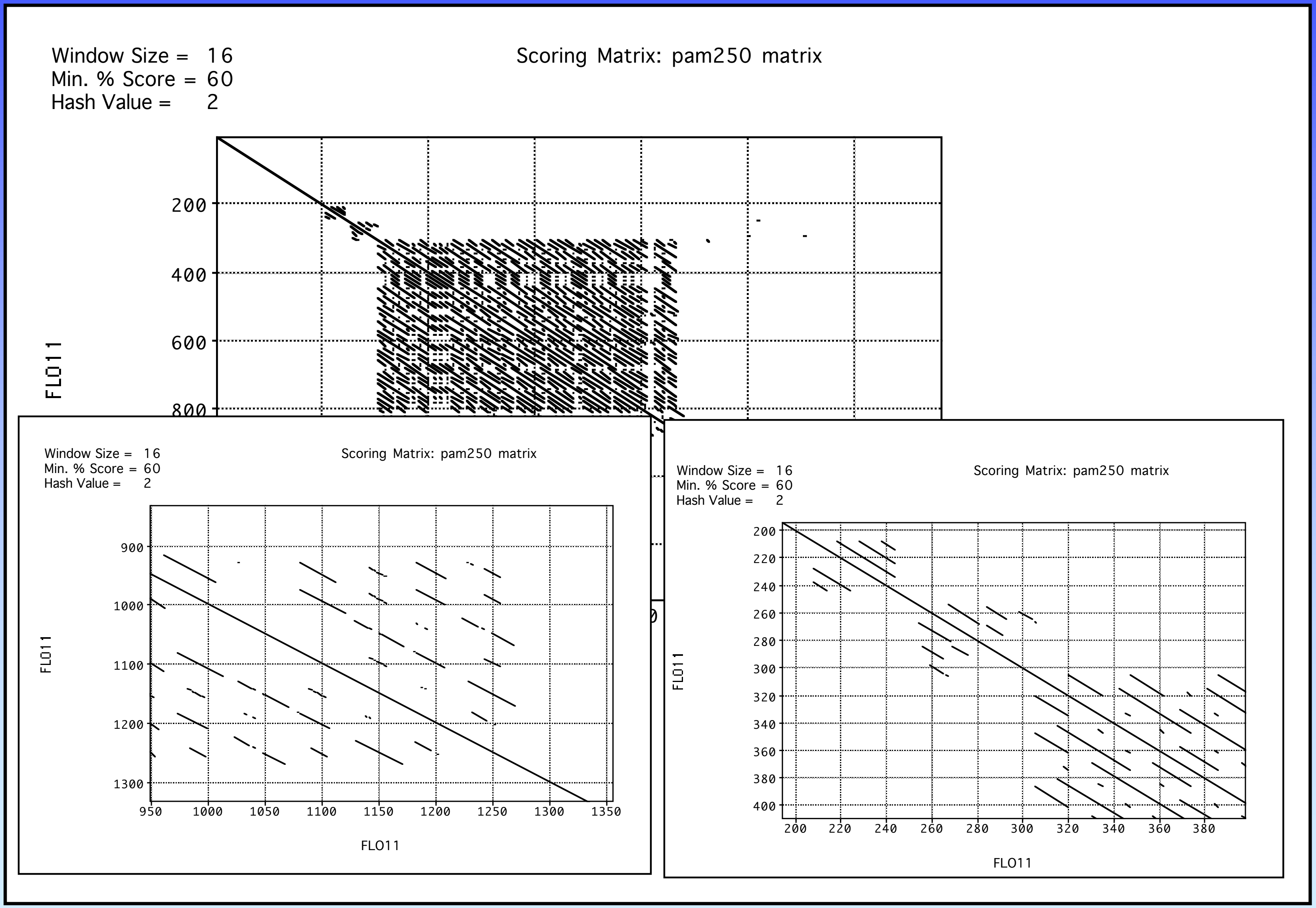
Dot Matrix Comparison



Window Size = 8
Min. % Score = 50
Hash Value = 2

Scoring Matrix: pam250 matrix

Dot Matrix Comparison



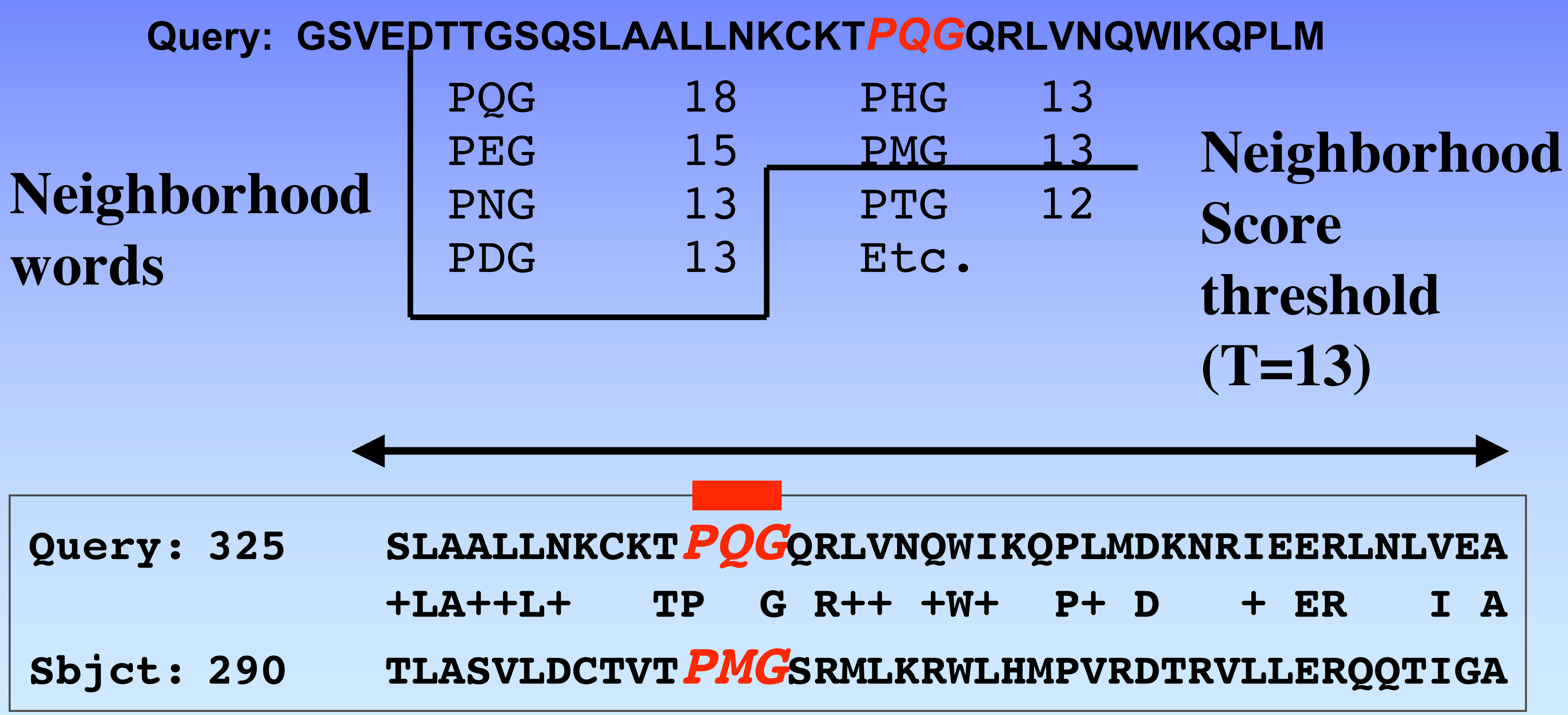
Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences
- Puts in gaps and mismatches
- Maximum number of matches between identical or related characters
- Generates a score and statistical assessment
- Nice example of global alignment using N-W: http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

BLAST Algorithm (1990) "Ungapped" alignment

- To improve speed, use a word based hashing scheme to index database
- Limit search for similarities to only the region near matching words
- Use **T**hreshold parameter to rate neighbor words
- Extend match left and right to search for high scoring alignments

Original BLAST Algorithm (1990)



Query word (W=3)

Query: GSVEDTTGSQSLAALLNKCKT*PQG*QRLVNQWIKQPLM

Neighborhood words

PQG	18	PHG	13
PEG	15	PMG	13
PNG	13	PTG	12
PDG	13	Etc.	

Neighborhood Score threshold (T=13)

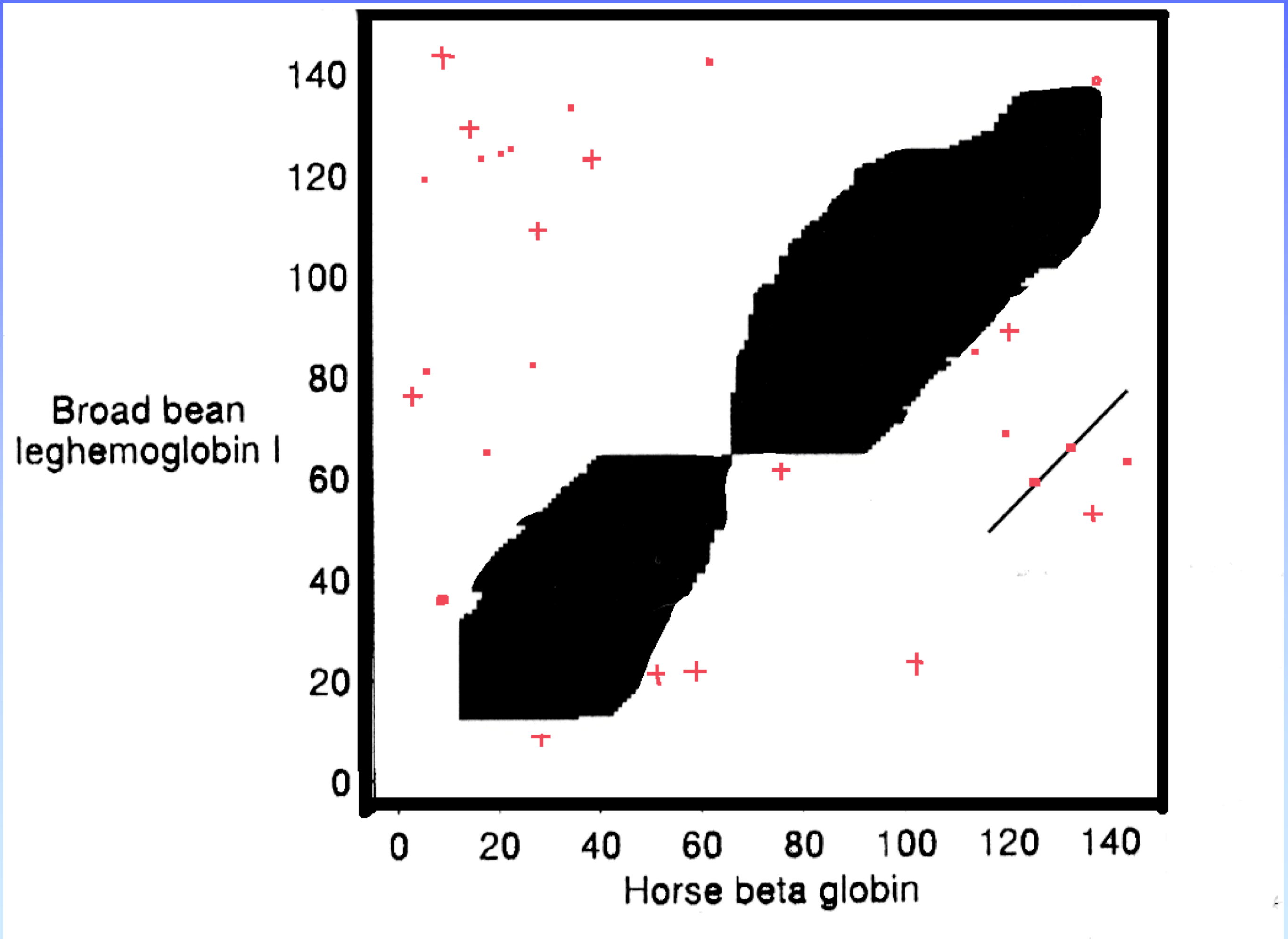
```
Query: 325   SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA
             +LA++L+   TP G R++ +W+  P+ D   + ER    I A
Sbjct: 290   TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA
```

BLAST Refinements (1997)

- “two-hit” method for extending word pairs
- Gapped alignments
- Additional algorithms
  - – Iterate with position-specific matrix (PSI-BLAST)
  - – Pattern-hit initiated BLAST (PHI-BLAST)

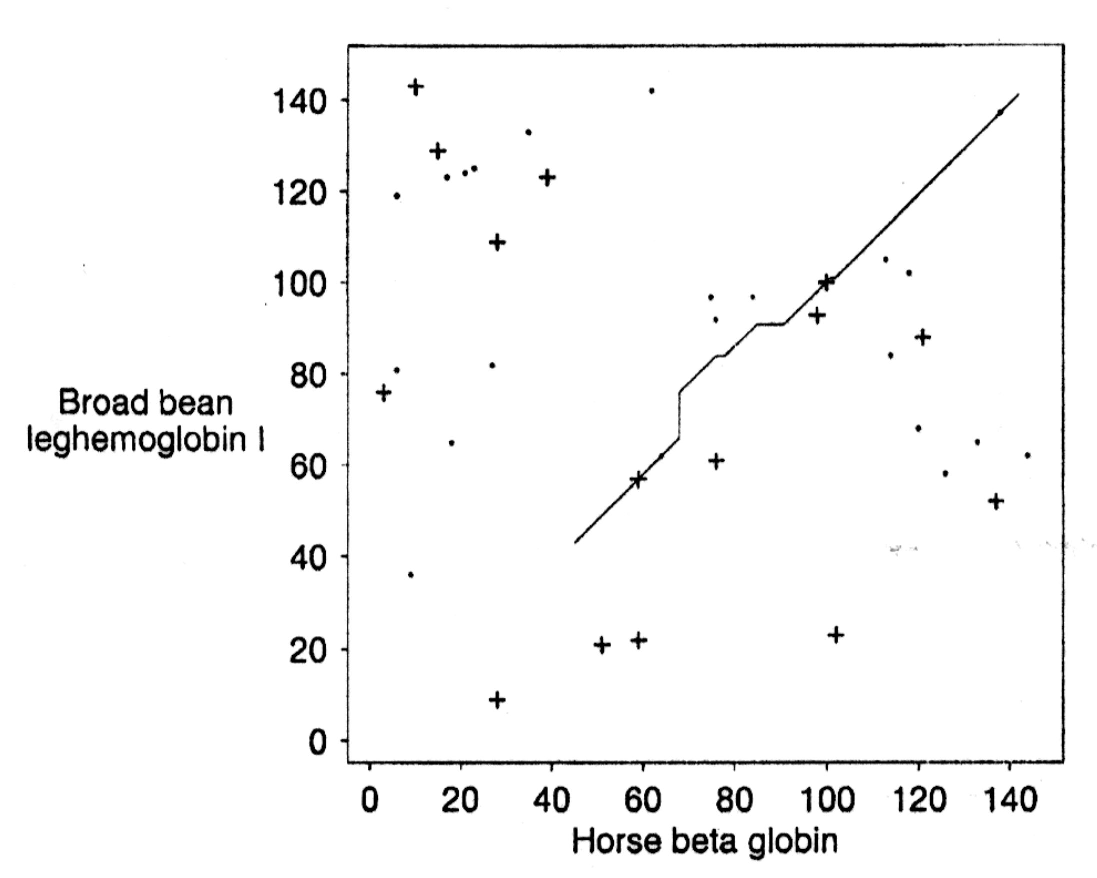
Gapped BLAST

15(+) > 13
22(•) > 11



(Altschul et al 1997)

Gapped BLAST



(Altschul et al 1997)

Programs to Compare two sequences - Unix or Web

NCBI

BLAST 2 Sequences

EMBOSS

water - Smith-Waterman

needle - Needleman -Wunsch

dotmatch (dot plot)

einverted or palindrome (inverted repeats)

equicktandem or etandem (tandem repeats)

Other

lalign (multiple matching subsegments in two sequences)

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
  - How strong can alignment be by chance alone?
- Database searching methods

Statistical Significance

- ***Raw Scores*** - score of an alignment equal to the sum of substitution and gap scores.
- ***Bit scores*** - scaled version of an alignment's raw score that accounts for the statistical properties of the scoring system used.
- ***E-value*** - expected number of distinct alignments that would achieve a given score by chance. Lower E-value => more significant.

Some formulas

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

This is the Expected number of high-scoring segment pairs (HSPs) with score at least S for sequences of length m and n.

This is the E value for the score S.

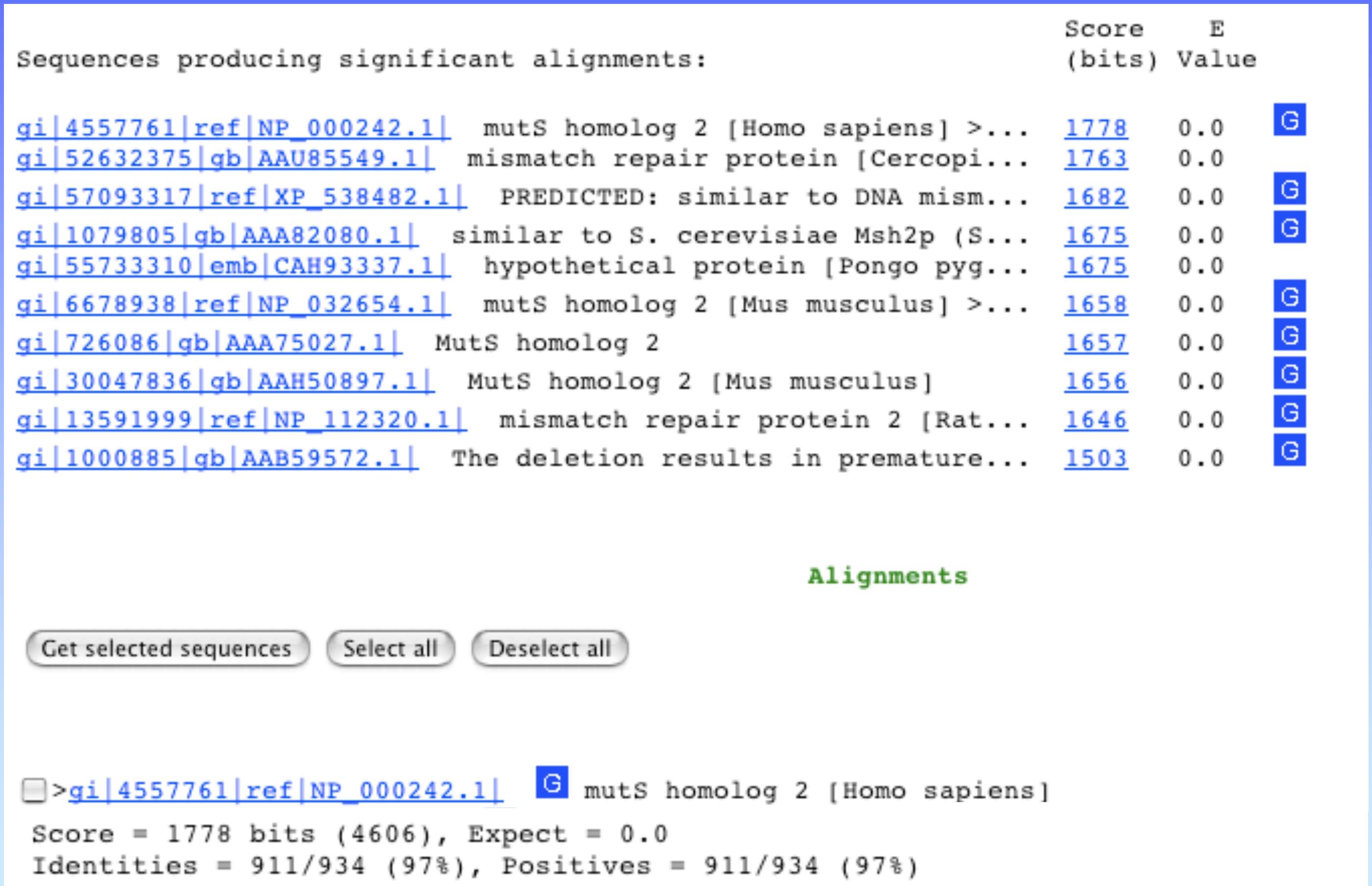
Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods
  - BLAST
  - BLAST vs. FASTA
  - BLAT

Questions

- Why do a database search?
- What database should be searched?
- What alignment algorithm to use?
- What do the results mean?
- What parameters can be changed?
  - Substitution matrices
  - Statistical significance
  - Filtering for low complexity

BLASTP Results



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

gi|4557761|ref|NP_000242.1|  mutS homolog 2 [Homo sapiens] >...   1778   0.0      G
gi|52632375|gb|AAU85549.1|  mismatch repair protein [Cercopi...   1763   0.0
gi|57093317|ref|XP_538482.1|  PREDICTED: similar to DNA mism...   1682   0.0      G
gi|1079805|gb|AAA82080.1|  similar to S. cerevisiae Msh2p (S...   1675   0.0      G
gi|55733310|emb|CAH93337.1|  hypothetical protein [Pongo pyg...   1675   0.0
gi|6678938|ref|NP_032654.1|  mutS homolog 2 [Mus musculus] >...   1658   0.0      G
gi|726086|gb|AAA75027.1|  MutS homolog 2                          1657   0.0      G
gi|30047836|gb|AAH50897.1|  MutS homolog 2 [Mus musculus]         1656   0.0      G
gi|13591999|ref|NP_112320.1|  mismatch repair protein 2 [Rat...   1646   0.0      G
gi|1000885|gb|AAB59572.1|  The deletion results in premature...   1503   0.0      G


                                              Alignments
```

```
□>gi|4557761|ref|NP_000242.1|  G  mutS homolog 2 [Homo sapiens]
 Score = 1778 bits (4606), Expect = 0.0
 Identities = 911/934 (97%), Positives = 911/934 (97%)
```

WU-BLAST vs NCBI BLAST

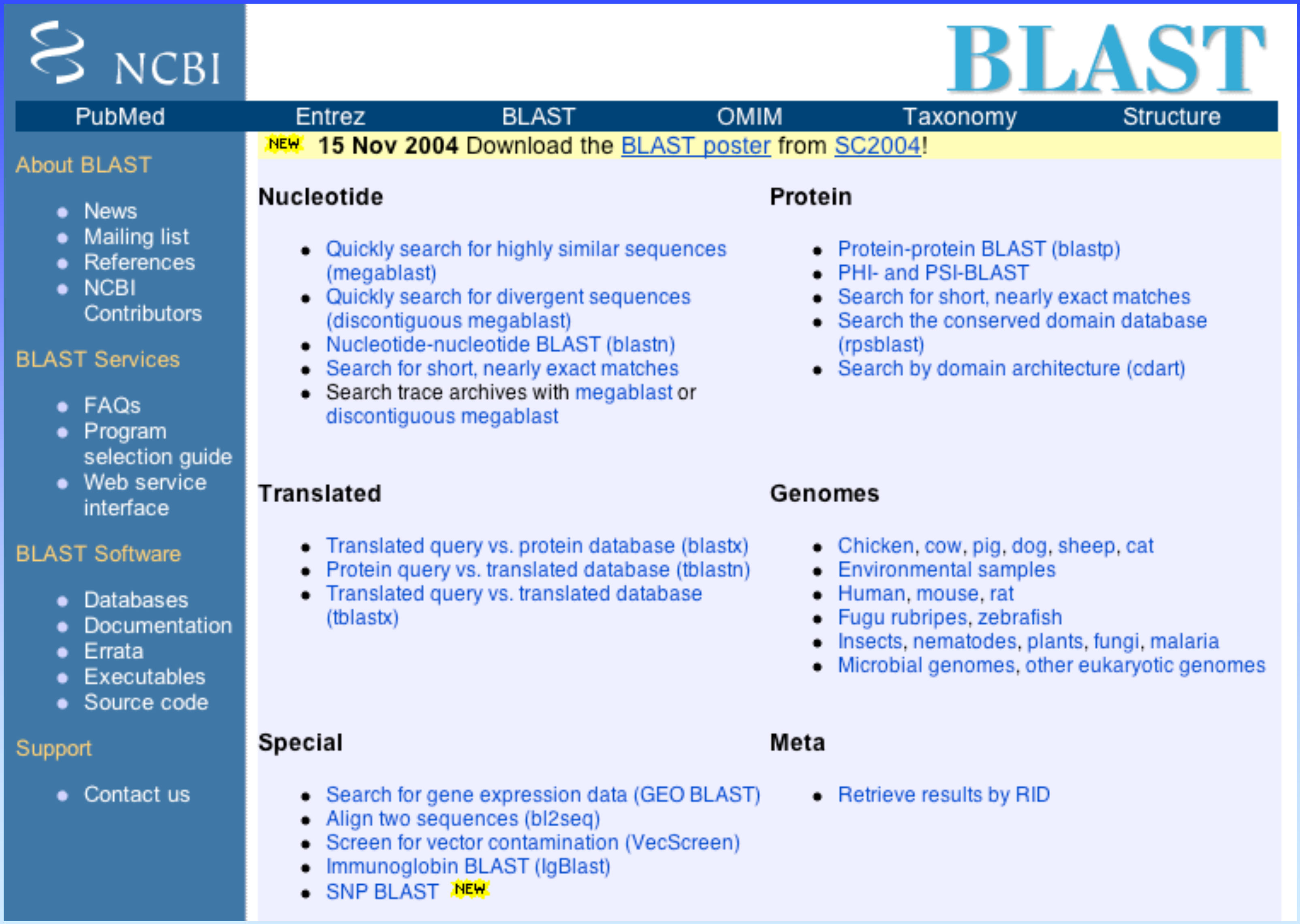
- WU-BLAST first for gapped alignments
- Use different scoring system for gaps
- Report different statistics
- WU-BLAST does not filter low-complexity by default
- WU-BLAST looks for and reports multiple regions of similarity
- Results will be different!

BLAT

- ***B***last-***L***ike ***A***lignment ***T***ool
- Developed by Jim Kent at UCSC
- For DNA it is designed to quickly find sequences of >= 95% similarity of length 40 bases or more.
- For proteins it finds sequences of >= 80% similarity of length 20 amino acids or more.
- DNA BLAT works by keeping an index of the entire genome in memory - non-overlapping 11-mers (< 1 GB of RAM)
- Protein BLAT uses 4-mers (~ 2 GB)

FASTA

- Index "words" and locate identities
- Rescore best 10 regions
- Find optimal subset of initial regions that can be joined to form single alignment
- Align highest scoring sequences using Smith-Waterman



NEW 15 Nov 2004 Download the BLAST poster from SC2004!

About BLAST

- News
- Mailing list
- References
- NCBI Contributors

BLAST Services

- FAQs
- Program selection guide
- Web service interface

BLAST Software

- Databases
- Documentation
- Errata
- Executables
- Source code

Support

- Contact us

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Chicken, cow, pig, dog, sheep, cat
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST NEW

Meta

- Retrieve results by RID

Basic Searching Strategies

- Search early and often
- Use specialized databases
- Use multiple matrices
- Use filters
- Consider Biology

Exercises

1. Working with Entrez at NCBI
   1. Limits, history, preview
   2. Batch Entrez
2. Pairwise alignment and database searching
   1. Self comparison with dottup
   2. Local vs. global alignment
   3. Sequence revision history, ReadSeq and BL2SEQ
   4. BLAT searching
   5. Comparing BLAST, WU-BLAST, and FASTA
   6. The Lost World
   7. Redo exercises on hebrides

References

1. Class web site:
   - http://jura.wi.mit.edu/bio/education/bioinfo2005/seq

2. Books
   - David Mount - Bioinformatics: Sequence and Genome Analysis, 2nd edition, CSHL Press, 2004.
   - Andreas D. Baxevanis and B. F. Francis Ouellette (editors) - Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd edition, Wiley, 2004.