Bioinformatics

Sequence Analysis: Part II. Database Searching

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute

Course Syllabus

Jan 7	Sequence Analysis I. Pairwise alignments, database searching including BLAST (FL) [1, 2, 3]
Jan 14	Sequence Analysis II. Database searching (continued), Pattern searching(FL)[7]
Jan 21	No Class - Martin Luther King Holiday
Jan 28	Sequence Analysis III. Hidden Markov models, gene finding algorithms (FL)[8]
Feb 4	Computational Methods I. Genomic Resources and Unix (GB)
Feb 11	Computational Methods II. Sequence analysis with Perl. (GB)
Feb 18	No Class - President's Birthday
Feb 25	Computational Methods III. Sequence analysis with Perl and BioPerl (GB)
Mar 4	Proteins I. Multiple sequence alignments, phylogenetic trees (RL) [4, 6]
Mar 11	Proteins II. Profile searches of databases, revealing protein motifs (RL) [9]
Mar 18	Proteins III.Structural Genomics:structural comparisons and predictions (RL)
Mar 25	Microarrays: designing chips, clustering methods (FL)

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)
- Significance of alignments
- Database searching methods
- Demo

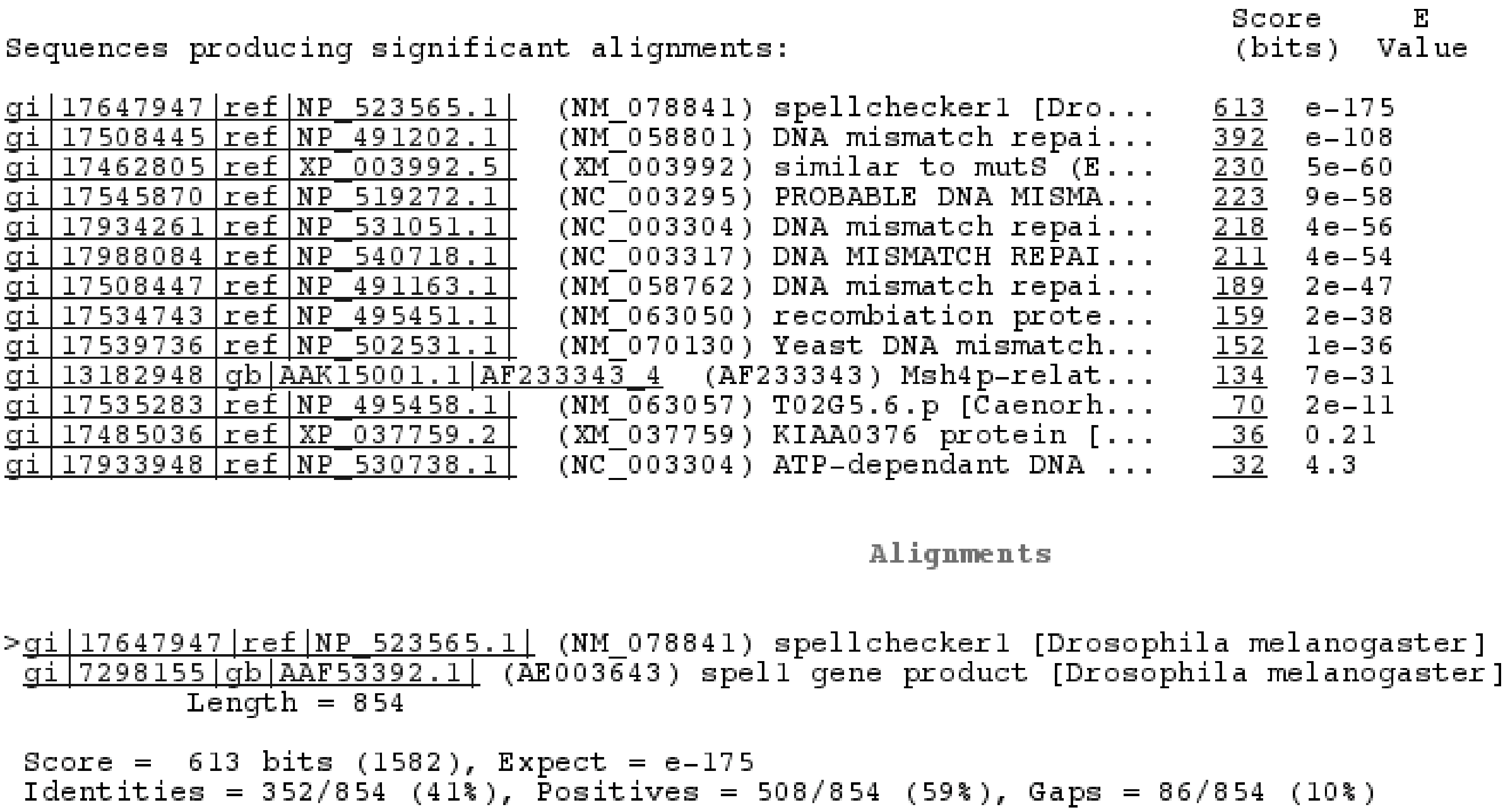
Questions

- Why do a database search?
- What database should be searched?
- What alignment algorithm to use?
- What do the results mean?

Issues affecting DB Search

- Substitution matrices
- Statistical significance
- Filtering
- Database choices

BLASTP Results



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

gi|17647947|ref|NP_523565.1|  (NM_078841) spellchecker1 [Dro...   613   e-175
gi|17508445|ref|NP_491202.1|  (NM_058801) DNA mismatch repai...   392   e-108
gi|17462805|ref|XP_003992.5|  (XM_003992) similar to mutS (E...   230   5e-60
gi|17545870|ref|NP_519272.1|  (NC_003295) PROBABLE DNA MISMA...   223   9e-58
gi|17934261|ref|NP_531051.1|  (NC_003304) DNA mismatch repai...   218   4e-56
gi|17988084|ref|NP_540718.1|  (NC_003317) DNA MISMATCH REPAI...   211   4e-54
gi|17508447|ref|NP_491163.1|  (NM_058762) DNA mismatch repai...   189   2e-47
gi|17534743|ref|NP_495451.1|  (NM_063050) recombiation prote...   159   2e-38
gi|17539736|ref|NP_502531.1|  (NM_070130) Yeast DNA mismatch...   152   1e-36
gi|13182948|gb|AAK15001.1|AF233343_4  (AF233343) Msh4p-relat...   134   7e-31
gi|17535283|ref|NP_495458.1|  (NM_063057) T02G5.6.p [Caenorh...    70   2e-11
gi|17485036|ref|XP_037759.2|  (XM_037759) KIAA0376 protein [...    36   0.21
gi|17933948|ref|NP_530738.1|  (NC_003304) ATP-dependant DNA ...    32   4.3


                                                  Alignments


>gi|17647947|ref|NP_523565.1| (NM_078841) spellchecker1 [Drosophila melanogaster]
 gi|7298155|gb|AAF53392.1| (AE003643) spell gene product [Drosophila melanogaster]
          Length = 854

 Score =  613 bits (1582), Expect = e-175
 Identities = 352/854 (41%), Positives = 508/854 (59%), Gaps = 86/854 (10%)
```

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)
- Significance of alignments
- Database searching methods
- Demo

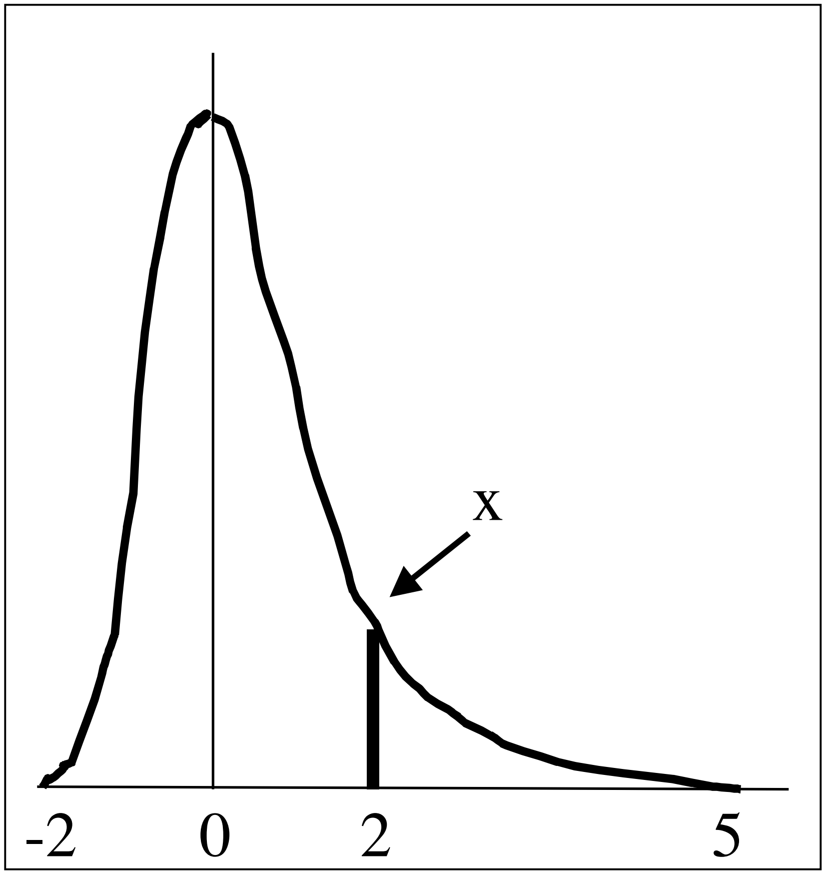
Significance of Alignment

How strong can an alignment be expected by chance alone?

- Real but non-homologous sequences
- Real sequences that are shuffled to preserve compositional properties
- Sequences that are generated randomly based upon a DNA or protein sequence model

Extreme Value Distribution

- When 2 sequences have been aligned optimally, the significance of a local alignment score can be tested on the basis of the distribution of scores expected by aligning two random sequences of the same length and composition as the two test sequences.



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 - ungapped and gapped
  - BLAST vs. FASTA
  - PSI-BLAST
  - PHI-BLAST
  - Pattern searching
- Demo

Low Complexity Regions

- **Local regions of biased composition**
- **Common in real sequences**
- **Generate false positives on BLAST search**

- **DUST for BLASTN (n’s in sequence)**
- **SEG for other programs (x’s in sequence)**

Filtering is only applied to the query sequence (or its translation products), not to database sequences.

Filtered Sequence

NEEYTKNKTEYEE

TALTTEETLT

```
>HUMAN MSH2
QGMPEKPTTTVRLFDRGDFYTAHGEDALLAAR
VLSKMNFESFVKDLLLVRQYRVEVYKNRAGNK
ILFGNNDMSASIGVVGVKMSAVDGQRQVGVGY
ALLIQIGPKECVLPGGETAGDMGKLRQIIQRG
GILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIK
FLELLSDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALL
NKCKTPQGQRLVNQWIKQPLMDKNRIEE
DLNRLAKKFQRQAANLQDCYRLYQGINQ
TDLRSDFSKFQEMIETTLDMDQVENHEF
LISAARDLGLDPGKQIKLDSSAQFGYYF
FTNSKLTSLNEEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAV
VSFAHVSNGAPVPYVRPAILEKGQGRIILKASRHACVEVQDEIAFIPNDVYFEKD
KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARVGAG
DSQLKGVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYI
ATKIGAFCMFATHFHELTALANQIPTVNNLHVTALTTEETLTMLYQVKKGVCDQS
FGIHVAELANFPKHVIECAKQKALELEEFQYIGESQGYDIMEPAAKKCYLEREQG
EKIIQEFLSKVKQMPFTEMSEENITIKLKQLKAEVIAKNNSFVNEIISRIKVTT
```

Example Alignment w/o filtering

Score = 29.6 bits (65), Expect = 1.8
Identities = 22/70 (31%), Positives = 32/70 (45%), Gaps = 12/70 (17%)

```
Query: 31   PPPTTQGAPRTSSFTPTTLT------------NGTSHSPTALNGAPSPPNGFS 71
            PPP+ Q   R  S +  T T            NG+S S ++ + + S  +  S
Sbjct: 1221 PPPSVQNQQRWGSSSVITTTCQQRQQSVSPHSNGSSSSSSSSSSSSSSSSSTS 1273

Query: 72   NGPSSSSSSSLANQQLP 88
            +  SSSS+SS    Q P
Sbjct: 1274 SNCSSSSASSCQYFQSP 1290
```

Example BLAST w/ filtering

```
 Score = 36.6 bits (83), Expect = 0.67
 Identities = 21/58 (36%), Positives = 25/58 (42%), Gaps = 1/58 (1%)

Query: 471 AEDALAVINQQEDSSESCWNCGRKASETCSGCNTARYCGSFCQHKDWE-KHHHICGQT 527
           A D   V  Q +  +  C  CG  A  TCS C  A YC    Q  DW+  H   C Q+
Sbjct: 61  ASDTECVCLQLKSGAHLCRVCGCLAPMTCSRCKQAHYCSKEHQTLDWQLGHKQACTQS 118
```

```
 Score = 37.0 bits (84), Expect = 0.55
 Identities = 18/55 (32%), Positives = 22/55 (39%)

Query: 483 DSSESCWNCGRKASETCSGCNTARYCGSFCQHKDWEKHHHICGQTLQAQQQGDTP 537
           D    C  CG  A++ C+ C  ARYC    Q  DW  H   C +        D P
Sbjct: 75  DGPGLCRICGCSAAKKCAKCQVARYCSQAHQVIDWPAHKLECAKAATDGSITDEP 129
```

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

- Developed by Jim Kent at UCSC
- BLAT is not BLAST
- 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)

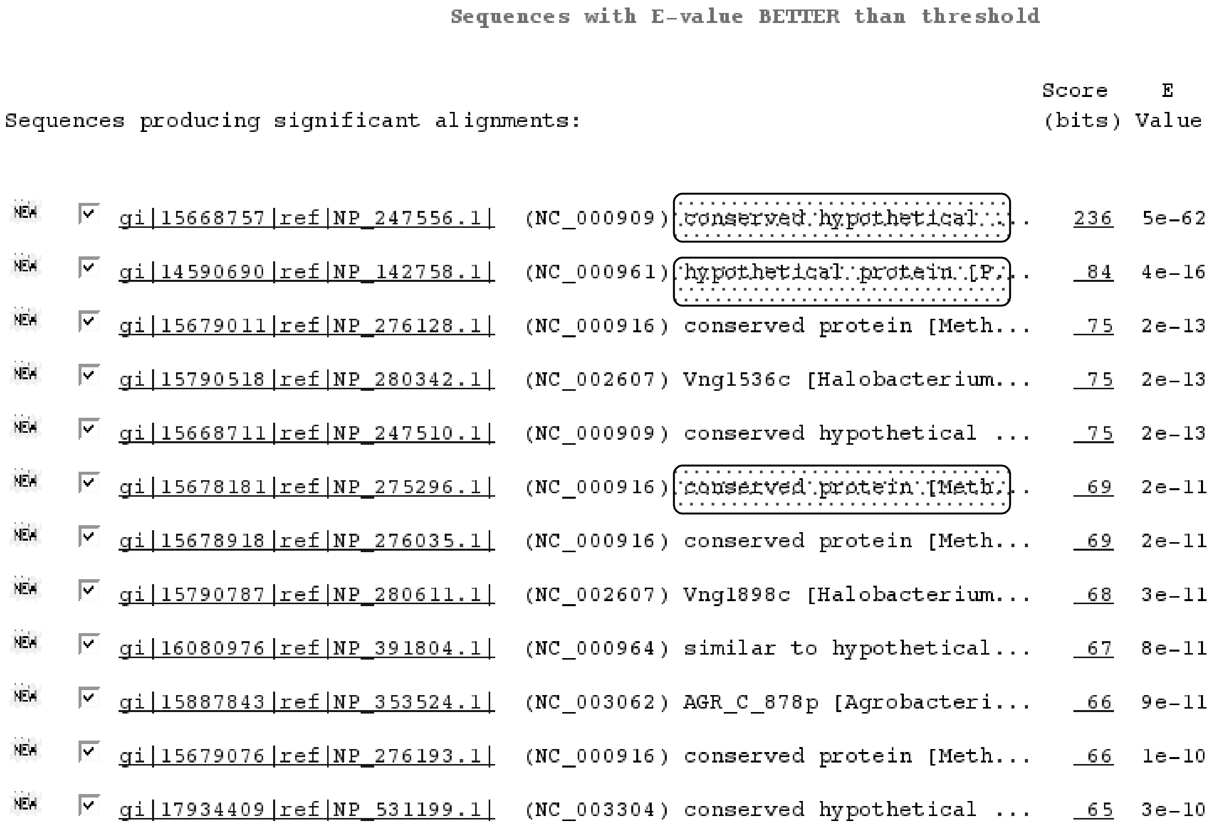
FASTA

- 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

PSI-BLAST

- Position specific iterative BLAST uses a profile (or position specific scoring matrix, PSSM) that is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search.
- The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero.
- The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" is used to refine the profile. This iterative searching strategy results in increased sensitivity.

Start with a BLASTP search

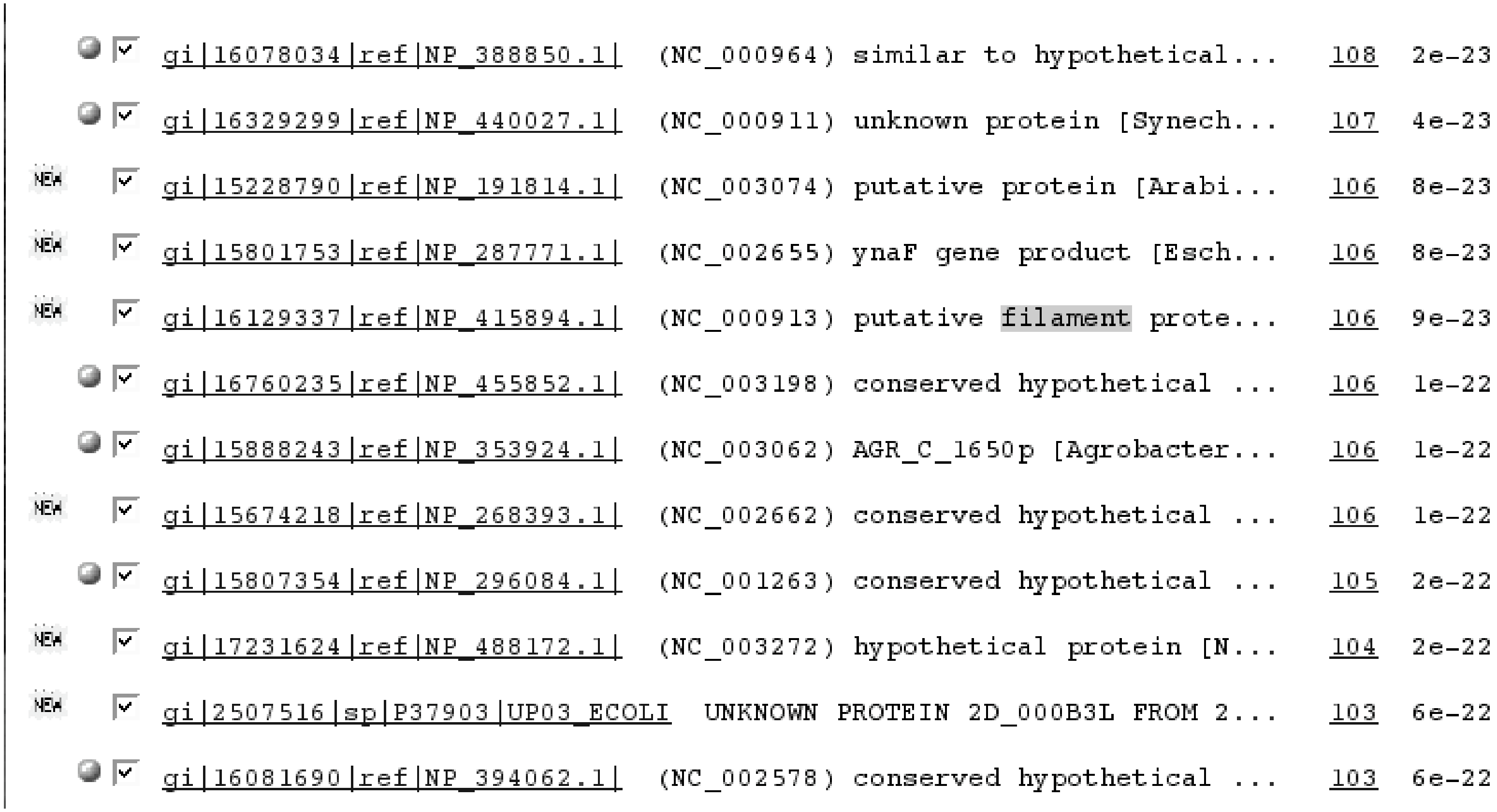


```
                          Sequences with E-value BETTER than threshold

                                                                                  Score    E
Sequences producing significant alignments:                                       (bits) Value

NEW ☑ gi|15668757|ref|NP_247556.1|  (NC_000909) conserved hypothetical ...   236  5e-62
NEW ☑ gi|14590690|ref|NP_142758.1|  (NC_000961) hypothetical protein [P...    84  4e-16
NEW ☑ gi|15679011|ref|NP_276128.1|  (NC_000916) conserved protein [Meth...    75  2e-13
NEW ☑ gi|15790518|ref|NP_280342.1|  (NC_002607) Vng1536c [Halobacterium...    75  2e-13
NEW ☑ gi|15668711|ref|NP_247510.1|  (NC_000909) conserved hypothetical ...    75  2e-13
NEW ☑ gi|15678181|ref|NP_275296.1|  (NC_000916) conserved protein [Meth...    69  2e-11
NEW ☑ gi|15678918|ref|NP_276035.1|  (NC_000916) conserved protein [Meth...    69  2e-11
NEW ☑ gi|15790787|ref|NP_280611.1|  (NC_002607) Vng1898c [Halobacterium...    68  3e-11
NEW ☑ gi|16080976|ref|NP_391804.1|  (NC_000964) similar to hypothetical...    67  8e-11
NEW ☑ gi|15887843|ref|NP_353524.1|  (NC_003062) AGR_C_878p [Agrobacteri...    66  9e-11
NEW ☑ gi|15679076|ref|NP_276193.1|  (NC_000916) conserved protein [Meth...    66  1e-10
NEW ☑ gi|17934409|ref|NP_531199.1|  (NC_003304) conserved hypothetical ...    65  3e-10
```

PSI-BLAST - Iteration 1



```
    ● ☑ gi|16078034|ref|NP_388850.1|  (NC_000964) similar to hypothetical...   108  2e-23
    ● ☑ gi|16329299|ref|NP_440027.1|  (NC_000911) unknown protein [Synech...   107  4e-23
NEW   ☑ gi|15228790|ref|NP_191814.1|  (NC_003074) putative protein [Arabi...   106  8e-23
NEW   ☑ gi|15801753|ref|NP_287771.1|  (NC_002655) ynaF gene product [Esch...   106  8e-23
NEW   ☑ gi|16129337|ref|NP_415894.1|  (NC_000913) putative filament prote...   106  9e-23
    ● ☑ gi|16760235|ref|NP_455852.1|  (NC_003198) conserved hypothetical ...   106  1e-22
    ● ☑ gi|15888243|ref|NP_353924.1|  (NC_003062) AGR_C_1650p [Agrobacter...   106  1e-22
NEW   ☑ gi|15674218|ref|NP_268393.1|  (NC_002662) conserved hypothetical ...   106  1e-22
    ● ☑ gi|15807354|ref|NP_296084.1|  (NC_001263) conserved hypothetical ...   105  2e-22
NEW   ☑ gi|17231624|ref|NP_488172.1|  (NC_003272) hypothetical protein [N...   104  2e-22
NEW   ☑ gi|2507516|sp|P37903|UP03_ECOLI  UNKNOWN PROTEIN 2D_000B3L FROM 2...   103  6e-22
    ● ☑ gi|16081690|ref|NP_394062.1|  (NC_002578) conserved hypothetical ...   103  6e-22
```

Amino acids

PSSM from PSI-BLAST

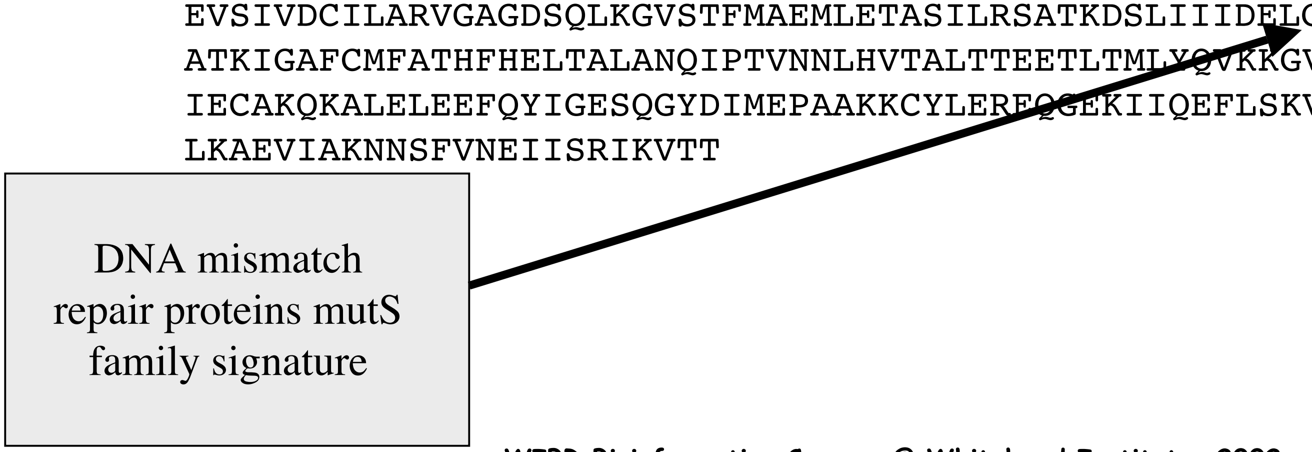
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0	2	3	2	4	1	1	4	3	0	3	3	7	3	3	2	1	0	1	2
2	6	0	3	3	5	4	0	3	2	5	0	1	2	2	4	1	3	2	4	2
3	4	3	0	3	3	1	3	2	4	2	3	2	5	0	1	2	1	0	5	7
4	3	2	3	2	4	9	3	3	5	4	0	3	2	5	1	2	2	4	1	2
5	0	1	2	2	4	1	6	3	3	1	3	2	0	4	8	3	1	0	3	0
6	4	3	2	…																
•	…																			
•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

POSITIONS

Pattern Hit Initiated (PHI)-BLAST

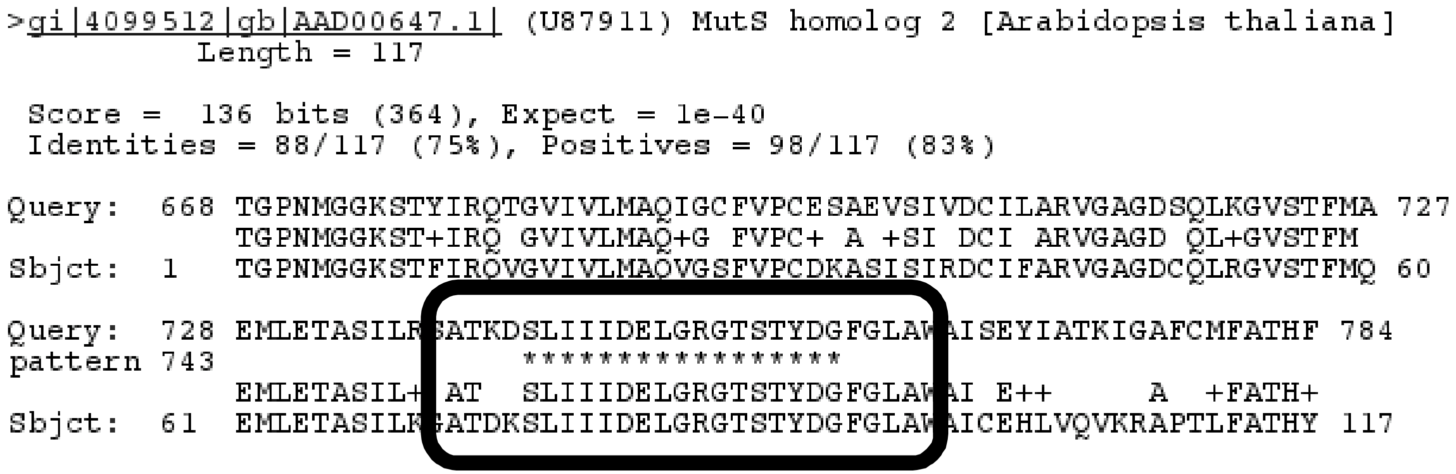
```
>HUMAN MSH2

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDALLAAREVFKTQGVIKYMGPA
GAKNLQSVVLSKMNFESFVKDLLLVRQYRVEVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNND
MSASIGVVGVKMSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECVLPGGETAGDM
GKLRQIIQRGGILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKFLELL
SDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPL
MDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAANLQDCYRLYQGINQLPNVIQA
LEKHEGKHQKLLLAVFVTPLTDLRSDFSKFQEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEK
KMQSTLISAARDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLN
EEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAILEKGQG
RIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA
EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYI
ATKIGAFCMFATHFHELTALANQIPTVNNLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHV
IECAKQKALELEEFQYIGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMSEENITIKLKQ
LKAEVIAKNNSFVNEIISRIKVTT
```



DNA mismatch repair proteins mutS family signature

PHI-BLAST



```
>gi|4099512|gb|AAD00647.1| (U87911) MutS homolog 2 [Arabidopsis thaliana]
          Length = 117

 Score =  136 bits (364), Expect = 1e-40
 Identities = 88/117 (75%), Positives = 98/117 (83%)

Query:  668 TGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMA 727
            TGPNMGGKST+IRQ GVIVLMAQ+G FVPC+ A +SI DCI ARVGAGD QL+GVSTFM
Sbjct:  1   TGPNMGGKSTFIRQVGVIVLMAQVGSFVPCDKASISIRDCIFARVGAGDCQLRGVSTFMQ 60

Query:  728 EMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHF 784
pattern 743                *****************
            EMLETASIL+ AT  SLIIIDELGRGTSTYDGFGLAWAI E++     A  +FATH+
Sbjct:  61  EMLETASILKGATDKSLIIIDELGRGTSTYDGFGLAWAICEHLVQVKRAPTLFATHY 117
```

Basic Searching Strategies

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

Demo

- Readseq
- Entrez
- NCBI
  - BLAST2.0
  - PSI-BLAST
  - PHI-BLAST
- WU-BLAST2
- FASTA
- Smith-Waterman

BLAST and FASTA Citations

- **PNAS, 1988, 85: 2444-2448.**
- **Journal of Molecular Biology, 1990, 215: 403-410.**
- **Nature Genetics, 1994, 6: 119-129.**
- **Nucleic Acids Research, 1997, 25(17):3389-3402.**
- **Nucleic Acids Research, 1998, 26(17):3986-3990.**
- **TIBS, 1998, 23:444-447.**
- **Nucleic Acids Research, 2001, 29(2):351-361.**
- **Nucleic Acids Research, 2001, 29(14):2994-3005.**

http://www.ncbi.nlm.nih.gov/blast
http://www.ebi.ac.uk/blast2/
http://www2.ebi.ac.uk/fasta33/
http://www2.ebi.ac.uk/bic_sw/