Advanced Database Searching

February 13, 2008

Jonathan Pevsner, Ph.D.
Introduction to Bioinformatics
Johns Hopkins University

Copyright notice

Outline of tonight's lecture

Specialized BLAST sites

Finding distantly related proteins: PSI-BLAST
PHI-BLAST

Profile Searches: Hidden Markov models

BLAST-like tools for genomic DNA
PatternHunter
Megablast
BLAT
BLASTZ

BLAST for gene discovery: Find-a-gene

Specialized BLAST servers

Species-specific BLAST sites

Molecule-specific BLAST sites

Specialized algorithms (WU-BLAST 2.0)

Page 135

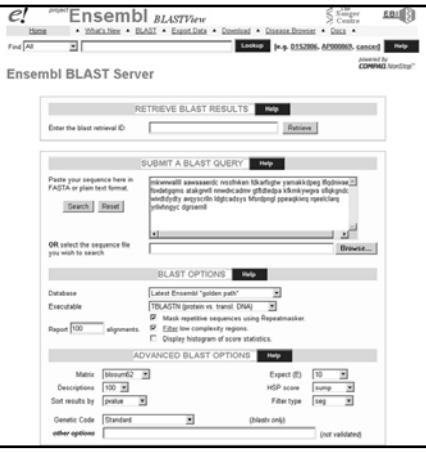
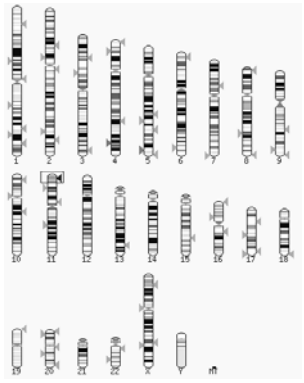
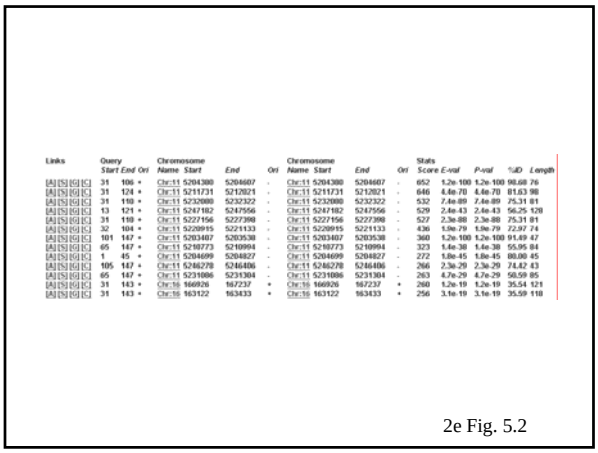


Fig. 5.1
Page 131

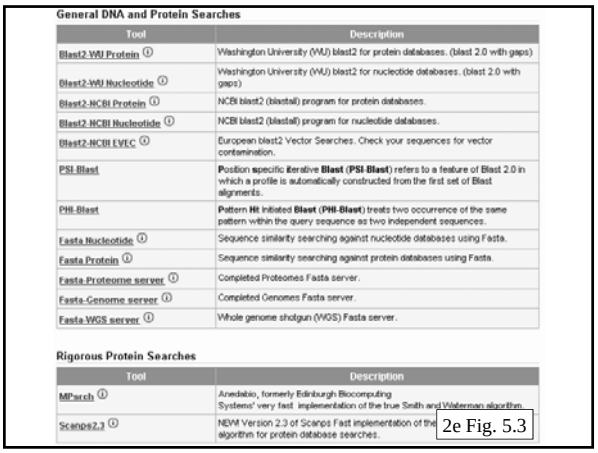
Ensembl BLAST output includes an ideogram



2e Fig. 5.1



2e Fig. 5.2



2e Fig. 5.3

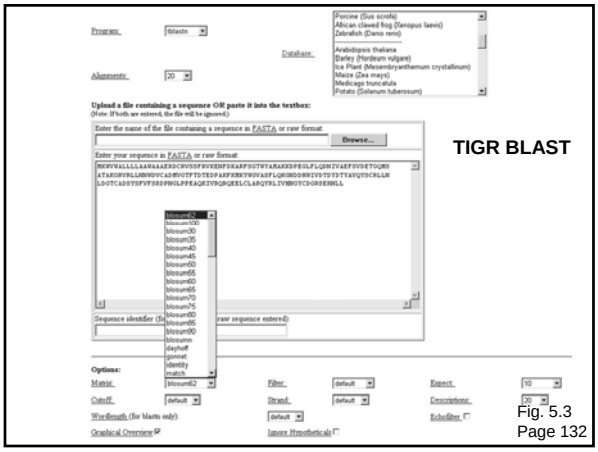


Fig. 5.3
Page 132

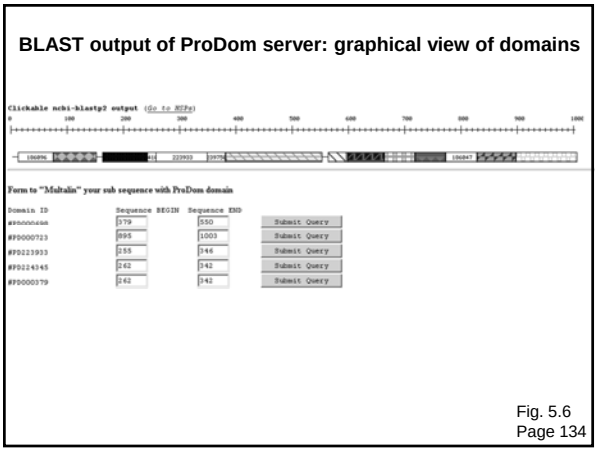


Fig. 5.6
Page 134

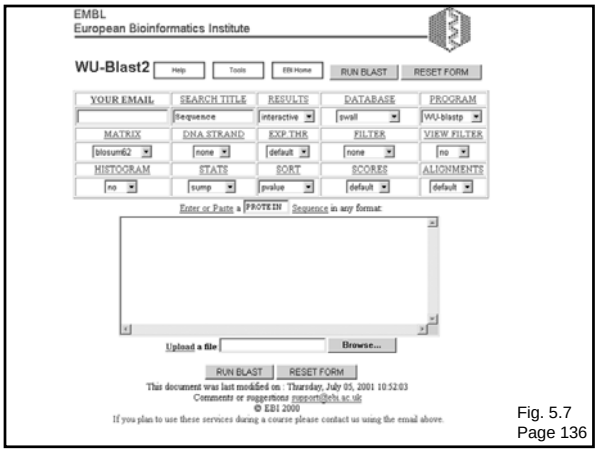


Fig. 5.7
Page 136

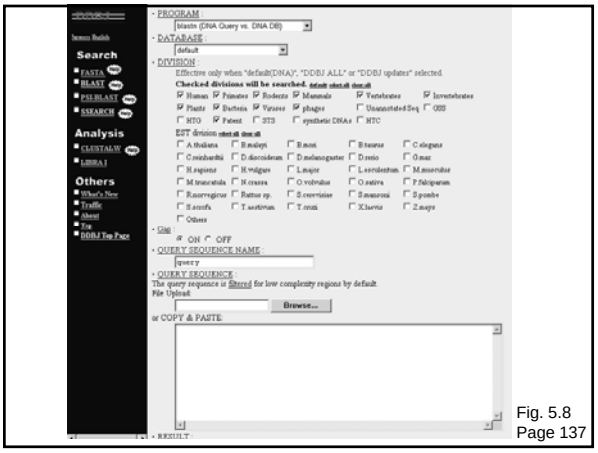
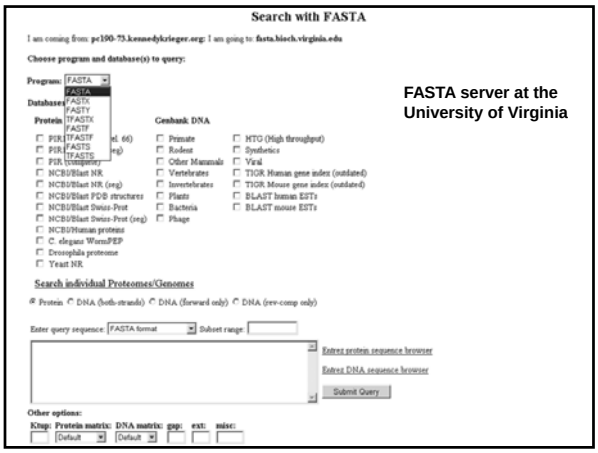


Fig. 5.8
Page 137



FASTA server at the University of Virginia

We will discuss the Conserved Domain Database (CDD) later (multiple sequence alignment)

Outline of today's lecture

Specialized BLAST sites

Finding distantly related proteins: PSI-BLAST
PHI-BLAST

Profile Searches: Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT
- BLASTZ

BLAST for gene discovery: Find-a-gene

Position specific iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

Page 137

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)

Page 138

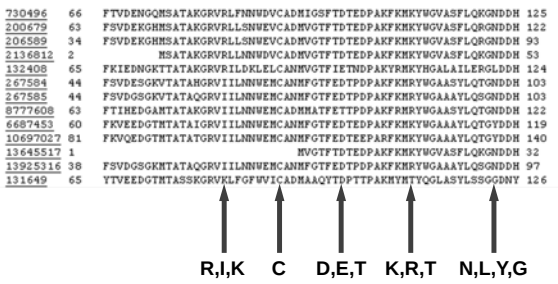
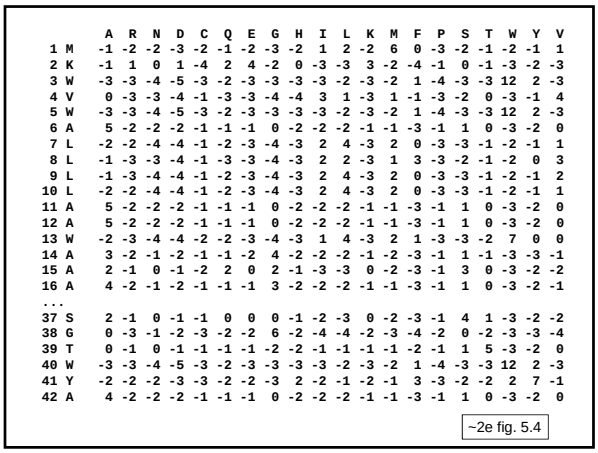
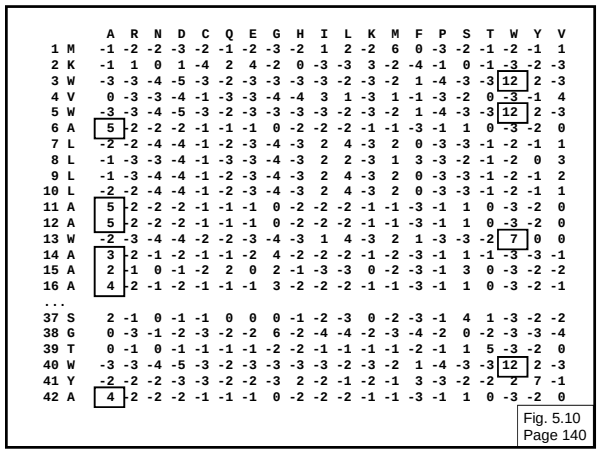


Fig. 5.9
Page 139



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

~2e fig. 5.4



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

Fig. 5.10
Page 140

PSI-BLAST is performed in five steps

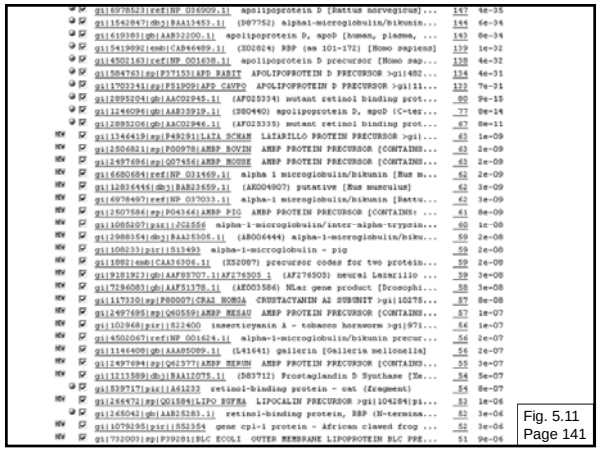
[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

Page 138



	gi|6978523|ref|NP_036909.1|	apolipoprotein D [Rattus norvegicus]...	147	4e-35
	gi|1542847|db||BAA13453.1|	(D87752) alpha1-microglobulin/bikunin...	144	6e-34
	gi|16191981|gb|AAH32200.1|	apolipoprotein D, apoD [Homo, plasma, ...	143	8e-34
	gi|5419892|emb|CAB46489.1|	(AD2024) RBP (aa 101-172) [Homo sapiens]	139	1e-32
	gi|4502163|ref|NP_001638.1|	apolipoprotein D precursor [Homo sap...	138	4e-32
	gi|584763|sp|P371531APD_RABIT	APOLIPOPROTEIN D PRECURSOR >gi|402...	134	4e-31
	gi|1702241|sp|P51909|APD_CAVPO	APOLIPOPROTEIN D PRECURSOR >gi|11...	123	7e-31
	gi|2895204|gb|AAC02945.1|	(AF025334) mutant retinol binding prot...	90	9e-15
	gi|1246096|gb|AAB23919.1|	(S80440) apolipoprotein D, apoD (C-ter...	77	8e-14
	gi|2895206|gb|AAC02946.1|	(AF025335) mutant retinol binding prot...	67	8e-11
HSP	gi|1346419|sp|P49291|LATA_SCHAM	LAZARILLO PROTEIN PRECURSOR >gi|...	63	1e-09
HSP	gi|2506621|sp|P00978|AMBP_BOVIN	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
HSP	gi|2497696|sp|Q07456|AMBP_MOUSE	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
HSP	gi|6680684|ref|NP_031469.1|	alpha 1 microglobulin/bikunin [Mus m...	62	2e-09
HSP	gi|12816446|db||BAB23659.1|	(AK004907) putative [Mus musculus]	62	3e-09
HSP	gi|6978497|ref|NP_037033.1|	alpha-1 microglobulin/bikunin [Rattu...	62	3e-09
HSP	gi|2507586|sp|P04366|AMBP_PIG	AMBP PROTEIN PRECURSOR [CONTAINS: ...	61	8e-09
HSP	gi|1085207|pir||JG2556	alpha-1-microglobulin/inter-alpha-trypsin...	60	1e-08
HSP	gi|2988354|db||BAA25306.1|	(AB006444) alpha-1-microglobulin/biku...	59	2e-08
HSP	gi|102233|pir||S13492	alpha-1-microglobulin - pig	59	2e-08
HSP	gi|1082|emb|CAA26906.1|	(X52007) precursor codes for two protein...	59	2e-09
HSP	gi|9181923|gb|AAF83707.1|AF176905_1	(AF176905) neural Lazarillo ...	59	3e-09
HSP	gi|7296083|gb|AAF51178.1|	(AE003586) HLaz gene product [Drosophi...	58	5e-08
HSP	gi|117330|sp|P80007|CRA2_HOMGA	CRUSTACYANIN A2 SUBUNIT >gi|10275...	57	8e-08
HSP	gi|2497695|sp|Q60559|AMBP_MESAU	AMBP PROTEIN PRECURSOR [CONTAINS...	57	1e-07
HSP	gi|102968|pir||S22400	insectocyanin A - tobacco hornworm >gi|971...	56	1e-07
HSP	gi|4502067|ref|NP_001624.1|	alpha-1-microglobulin/bikunin precur...	56	2e-07
HSP	gi|1146400|gb|AAA85089.1|	(L41641) gallerin [Galleria mellonella]	56	2e-07
HSP	gi|2497694|sp|Q62577|AMBP_MESUN	AMBP PROTEIN PRECURSOR [CONTAINS...	55	3e-07
HSP	gi|1213589|db||BAA12075.1|	(D83712) Prostaglandin D Synthase [Ra...	54	5e-07
	gi|539717|pir||A61232	retinol-binding protein - oat (fragment)	54	8e-07
HSP	gi|266471|sp|Q01584|LIPO_BOVBA	LIPOCALIN PRECURSOR >gi|104284|pi...	53	1e-06
	gi|2680542|gb|AAB25283.1|	retinol-binding protein, RBP (N-termina...	52	3e-06
HSP	gi|1079295|pir||S52354	gene cpl-1 protein - African clawed frog ...	52	3e-06
HSP	gi|732003|sp|P39281|BLC_ECOLI	OUTER MEMBRANE LIPOPROTEIN BLC PRE...	51	9e-06

Fig. 5.11
Page 141

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

[5] Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new profile is used as the query.

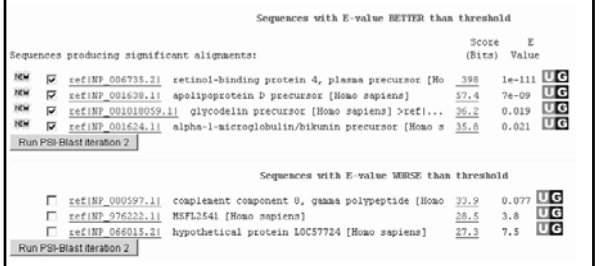
Page 138

Results of a PSI-BLAST search

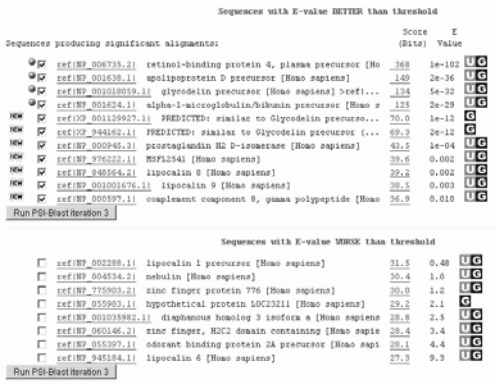
Iteration	# hits	# hits > threshold
1	104	49
2	173	96
3	236	178
4	301	240
5	344	283
6	342	298
7	378	310
8	382	320

Table 5-4
Page 141

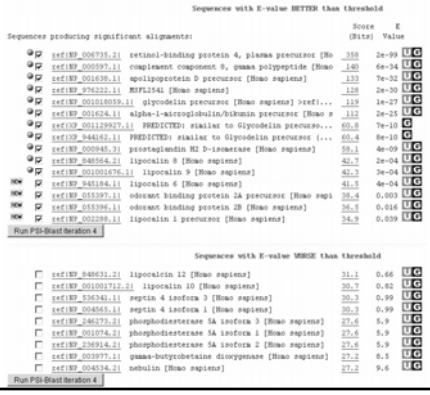
PSI-BLAST search: human RBP versus RefSeq, iteration 1



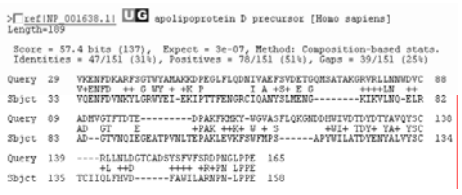
PSI-BLAST search: human RBP versus RefSeq, iteration 2



PSI-BLAST search: human RBP versus RefSeq, iteration 3

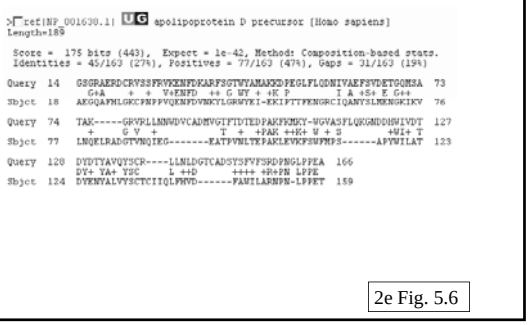


RBP4 match to ApoD, PSI-BLAST iteration 1



2e Fig. 5.6

RBP4 match to ApoD, PSI-BLAST iteration 2

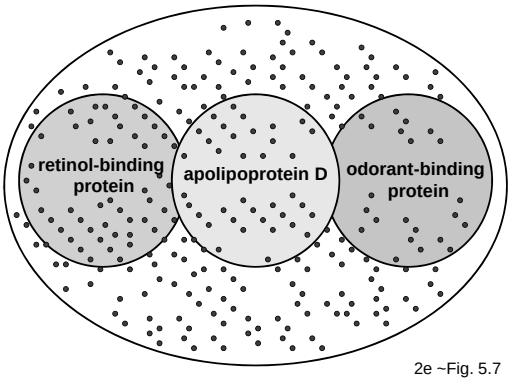


2e Fig. 5.6

RBP4 match to ApoD, PSI-BLAST iteration 3

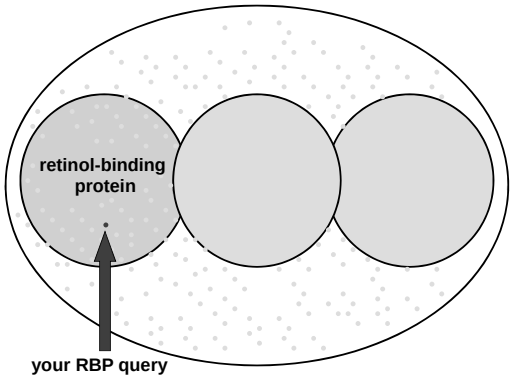
2e Fig. 5.6

The universe of lipocalins (each dot is a protein)

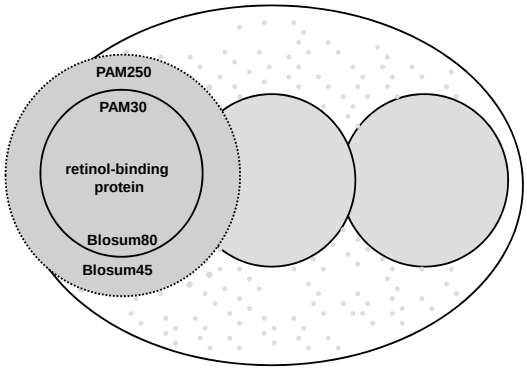


2e ~Fig. 5.7

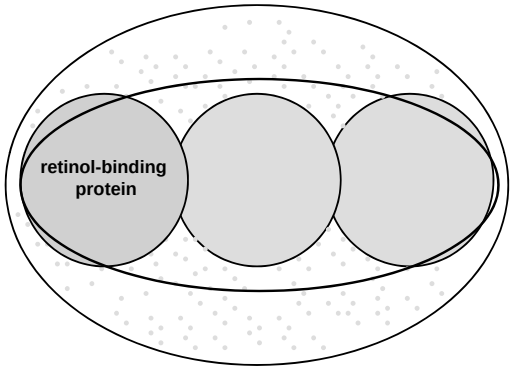
Scoring matrices let you focus on the big (or small) picture



Scoring matrices let you focus on the big (or small) picture



PSI-BLAST generates scoring matrices more powerful than PAM or BLOSUM



PSI-BLAST: performance assessment

Evaluate PSI-BLAST results using a database in which protein structures have been solved and all proteins in a group share $\leq$ 40% amino acid identity.

Page 143

PSI-BLAST: the problem of corruption

PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.

The main source of false positives is the spurious amplification of sequences not related to the query. For instance, a query with a coiled-coil motif may detect thousands of other proteins with this motif that are not homologous.

Once even a single spurious protein is included in a PSI-BLAST search above threshold, it will not go away.

Page 144

PSI-BLAST: the problem of corruption

Corruption is defined as the presence of at least one false positive alignment with an E value < 10^{-4} after five iterations.

Three approaches to stopping corruption:

[1] Apply filtering of biased composition regions

[2] Adjust E value from 0.001 (default) to a lower value such as E = 0.0001.

[3] Visually inspect the output from each iteration. Remove suspicious hits by unchecking the box.

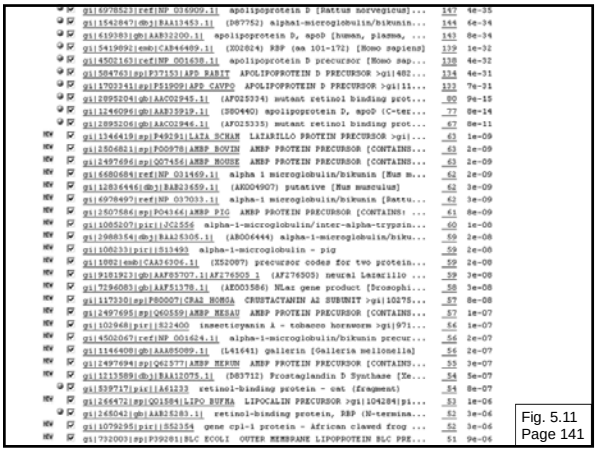


Fig. 5.11
Page 141

Conserved domain database (CDD) uses RPS-BLAST

2e Fig. 5.

PHI-BLAST: Pattern hit initiated BLAST

Launches from the same page as PSI-BLAST

Combines matching of regular expressions with local alignments surrounding the match.

PHI-BLAST: Pattern hit intiated BLAST

Launches from the same page as PSI-BLAST

Combines matching of regular expressions with local alignments surrounding the match.

Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question: What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences? PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology.

Align three lipocalins (RBP and two bacterial lipocalins)

```
      1                                                  50
ecblc  MRLLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDAKRYL GTWYEIARFD
  vc   MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWYEVARLD
hsrbp  ~~~MKWVWAL LLLAAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD
```

Fig. 5.9

Pick a small, conserved region and see which amino acid residues are used

```
       1                                                    50
ecblc  MRLLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDAKRYL GTWYEIARFD
   vc  MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWYEVARLD
hsrbp  ~~~MKWVWAL LLLAAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD

                                                  GTWYEI
                                                   K  AV
                                                      M
```

Fig. 5.9

Create a pattern using the appropriate syntax

```
       1                                                    50
ecblc  MRLLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDAKRYL GTWYEIARFD
   vc  MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWYEVARLD
hsrbp  ~~~MKWVWAL LLLAAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD

                                                  GTWYEI
                                                   K  AV
                                                      M
```

GXW[YF][EA][IVLM]

Fig. 5.9

Syntax rules for PHI-BLAST

The syntax for patterns in PHI-BLAST follows the conventions of PROSITE (protein lecture, Chapter 8).

When using the stand-alone program, it is permissible to have multiple patterns. When using the Web-page only one pattern is allowed per query.

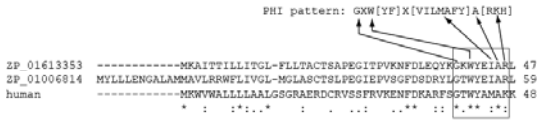
[ ] means any one of the characters enclosed in the brackets e.g., [LFYT] means one occurrence of L or F or Y or T

- means nothing (spacer character)

x(5) means 5 positions in which any residue is allowed

x(2,4) means 2 to 4 positions where any residue is allowed

PHI-BLAST

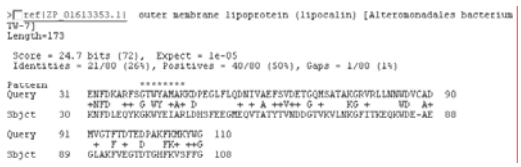


2e Fig. 5.9

PHI-BLAST: input

2e Fig. 5.9

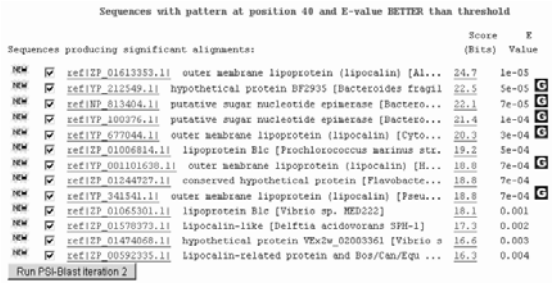
PHI-BLAST: output



2e Fig. 5.10

Human RBP PSI-BLAST search against bacteria

PHI-BLAST: RBP search against bacteria



2e Fig. 5.10

Outline of today's lecture

Specialized BLAST sites

Finding distantly related proteins: PSI-BLAST
PHI-BLAST

Profile Searches: Hidden Markov models

BLAST-like tools for genomic DNA
PatternHunter
Megablast
BLAT
BLASTZ

BLAST for gene discovery: Find-a-gene

Multiple sequence alignment to profile HMMs

• in 90's people began to see that aligning sequences to profiles gave much more information than pairwise alignment alone.

• Hidden Markov models (HMMs) are "states" that describe the probability of having a particular amino acid residue at arranged In a column of a multiple sequence alignment

• HMMs are probabilistic models

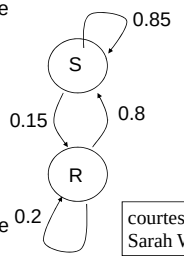
• Like a hammer is more refined than a blast, an HMM gives more sensitive alignments than traditional techniques such as progressive alignments

Page 325

Simple Markov Model

Rain = dog may not want to go outside

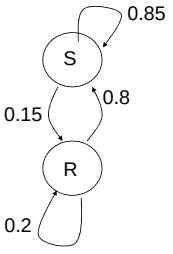
Sun = dog will probably go outside



Markov condition = no dependency on anything but nearest previous state ("memoryless")

courtesy of Sarah Wheelan

Simple Hidden Markov Model



P(dog goes out in rain) = 0.1

P(dog goes out in sun) = 0.85

Observation: YNNNYYNNNYN

(Y=goes out, N=doesn't go out)

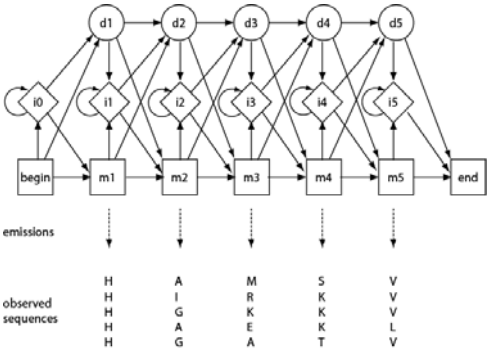
What is underlying reality (the hidden state chain)?

		Probability	position 1	2	3	4	5
		p(H)	1.0				
		p(A)		0.4			
		p(I)		0.2			
1D8U	HAMSV	p(G)		0.4			
1OJ6A	HIRKV	p(M)			0.2		
2hhbB	HGKKV	p(R)			0.2		
1FSL	HAEKL	p(K)			0.2		
2MM1	HGATV	p(E)			0.2		
		p(A)			0.2		
		p(S)				0.2	
		p(K)				0.6	
		p(T)				0.2	
		p(V)					0.8
		p(L)					0.2

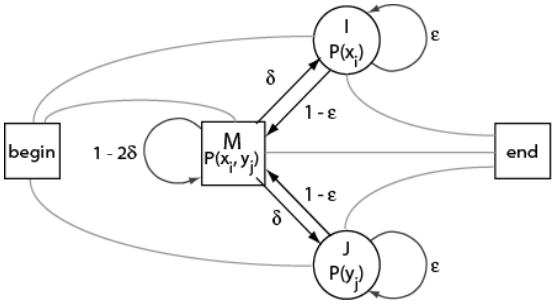
2e Fig. 5.11

p(HARTV) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128
Log odds score = ln(1.0) + ln(0.4) + ln(0.2) + ln(0.2) + ln(0.8) =

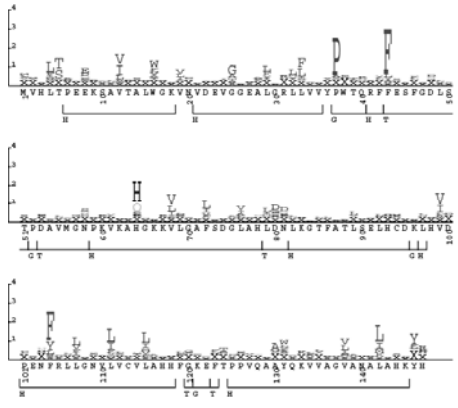
2e Fig. 5.11



2e Fig. 5.12



2e Fig. 5.12



HMMER: build a hidden Markov model

Determining effective sequence number ... done. [4]
Weighting sequences heuristically ... done.
Constructing model architecture ... done.
Converting counts to probabilities ... done.
Setting model name, etc. ... done. [x]

Constructed a profile HMM (length 230)
Average score: 411.45 bits
Minimum score: 353.73 bits
Maximum score: 460.63 bits
Std. deviation: 52.58 bits

Fig. ~5.13

HMMER: calibrate a hidden Markov model

HMM file: lipocalins.hmm
Length distribution mean: 325
Length distribution s.d.: 200
Number of samples: 5000
random seed: 1034351005
histogram(s) saved to: [not saved]
POSIX threads: 2
- -

HMM : x
mu : -123.894508
lambda : 0.179608
max : -79.334000

HMMER: search an HMM against GenBank

```
Scores for complete sequences (score includes all domains):
Sequence                          Description       Score    E-value  N
--------                          -----------       -----    -------  ---
gi|20888903|ref|XP_129259.1|      (XM_129259) ret   461.1   1.9e-133   1
gi|132407|sp|P04916|RETB_RAT      Plasma retinol-   458.0   1.7e-132   1
gi|20548126|ref|XP_005907.5|      (XM_005907) sim   454.9   1.4e-131   1
gi|5803139|ref|NP_006735.1|       (NM_006744) ret   454.6   1.7e-131   1
gi|20141667|sp|P02753|RETB_HUMAN  Plasma retinol-   451.1   1.9e-130   1
.
.
.
gi|16767588|ref|NP_463203.1|      (NC_003197) out   318.2    1.9e-90   1



gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 195: score 454.6, E = 1.7e-131
                  *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                     mkwV++LLLLaA  +  +aAErd      Crv+s    frVkeNFD+
  gi|5803139     1    MKWVWALLLLAA--W--AAAERD------CRVSS----FRVKENFDK 33

                     erylGtWYeIaKkDprFErGLllqdkItAeySleEhGsMsataeGrirVL
                     +r++GtWY++aKkDp  E GL+lqd+I+Ae+S++E+G+Msata+Gr+r+L
  gi|5803139    34 ARFSGTWYAMAKKDP--E-GLFLQDNIVAEFSVDETGQMSATAKGRVRLL 80

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                     +N+++cAD+vGT+t++E          dPak+k+Ky+GvaSflq+G+dd+
  gi|5803139    81 NNWDVCADMVGTFTDTE----------DPAKFKMKYWGVASFLQKGNDDH 120
```

HMMER: search an HMM against GenBank
match to a bacterial lipocalin

```
gi|16767588|ref|NP_463203.1|: domain 1 of 1, from 1 to 177: score 318.2, E = 1.9e-90
                  *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                        M+LL+ +A      a ++ Af+v++C++p+PP+G++V++NFD+
  gi|1676758     1    ----MRLLPVVA------AVTA-AFLVVACSSPTPPKGVTVVNNFDA 36

                     erylGtWYeIaKkDprFErGLllqdkItAeySleEhGsMsataeGrirVL
                     +rylGtWYeIa+ D+rFErGL   + +tA+ySl++          +G+i+V+
  gi|1676758    37 KRYLGTWYEIARLDHRFERGL---EQVTATYSLRD--------DGGINVI 75

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                      Nk++++D+  +++ +EG+a   ++t+ P +++lK+    Sf++p++++y
  gi|1676758    76 -NKGYNPDR-EMWQKTEGKA---YFTGSPNRAALKV----SFFGPFYGGY 116
```

HMMER: search an HMM against GenBank

```
Scores for complete sequences (score includes all domains):
Sequence                          Description       Score    E-value  N
--------                          -----------       -----    -------  ---
gi|3041715|sp|P27485|RETB_PIG      Plasma retinol-   614.2   1.6e-179   1
gi|89271|pir||A39486              plasma retinol-   613.9   1.9e-179   1
gi|20888903|ref|XP_129259.1|      (XM_129259) ret   608.8   6.8e-178   1
gi|132407|sp|P04916|RETB_RAT      Plasma retinol-   608.0   1.1e-177   1
gi|20548126|ref|XP_005907.5|      (XM_005907) sim   607.3   1.9e-177   1
gi|20141667|sp|P02753|RETB_HUMAN  Plasma retinol-   605.3   7.2e-177   1
gi|5803139|ref|NP_006735.1|       (NM_006744) ret   600.2   2.6e-175   1



gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 199: score 600.2, E = 2.6e-175
                  *->meWVWaLvLLaalGgasaERDCRvssFRVKEnFDKARFsGtWYAiAK
                     m+WVWaL+LLaa+  a+aERDCRvssFRVKEnFDKARFsGtWYA+AK
  gi|5803139     1    MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAK 45

                     KDPEGLFLqDnivAEFsvDEkGhmsAtAKGRvRLLnnWdvCADmvGtFtD
                     KDPEGLFLqDnivAEFsvDE+G+msAtAKGRvRLLnnWdvCADmvGtFtD
  gi|5803139    46 KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTD 95

                     tEDPAKFKmKYWGvAsFLqkGnDDHWiiDtDYdtfAvqYsCRLlnLDGtC
                     tEDPAKFKmKYWGvAsFLqkGnDDHWi+DtDYdt+AvqYsCRLlnLDGtC
  gi|5803139    96 TEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTC 145
```

Outline of today's lecture

Specialized BLAST sites

Finding distantly related proteins: PSI-BLAST
PHI-BLAST

Profile Searches: Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT
- BLASTZ

BLAST for gene discovery: Find-a-gene

BLAST-related tools for genomic DNA

The analysis of genomic DNA presents special challenges:
- There are exons (protein-coding sequence) and introns (intervening sequences).
- There may be sequencing errors or polymorphisms
- The comparison may between be related species (e.g. human and mouse)

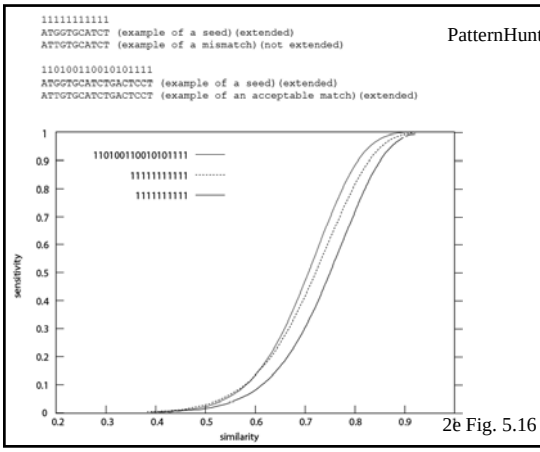
BLAST-related tools for genomic DNA

Recently developed tools include:

- MegaBLAST at NCBI.
- BLAT (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into words (11mers), then searches them against a query. Thus it is a mirror image of the BLAST strategy. See http://genome.ucsc.edu
- SSAHA at Ensembl uses a similar strategy as BLAT. See http://www.ensembl.org

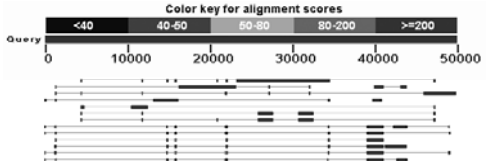
Page 136

PatternHunter

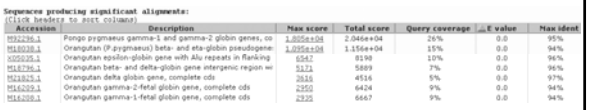


2e Fig. 5.16

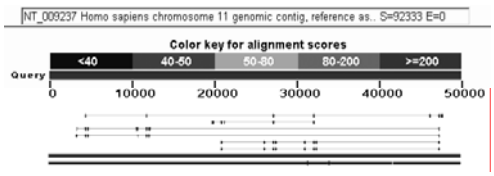
MegaBLAST at NCBI



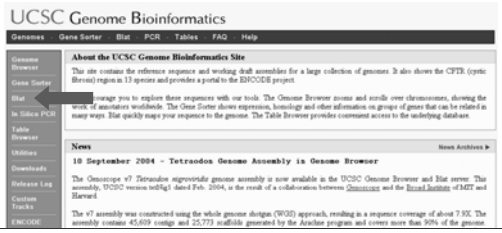
MegaBLAST



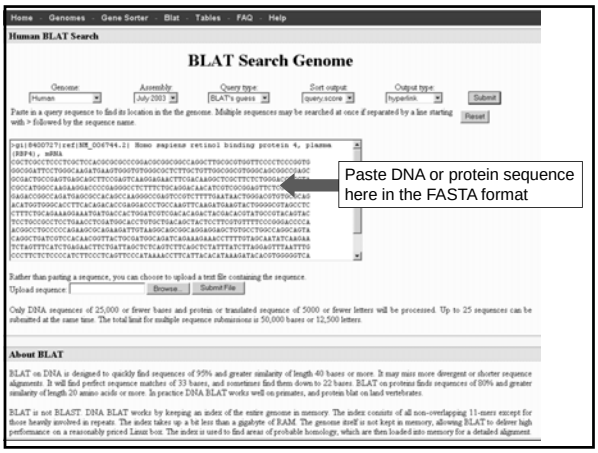
MegaBLAST: 50 kilobases of the globin locus



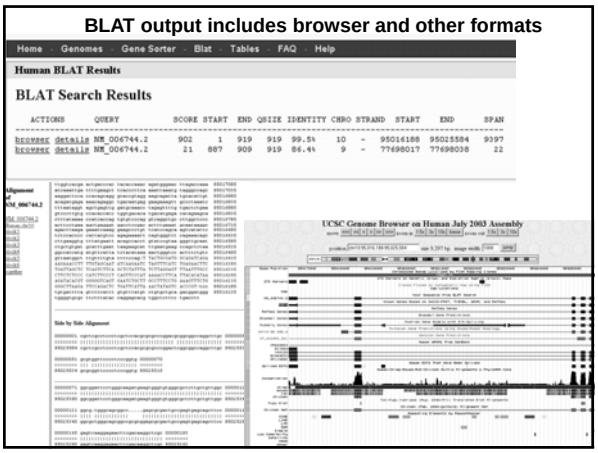
To access BLAT, visit http://genome.ucsc.edu



"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates." --BLAT website



BLAT output includes browser and other formats



Blastz

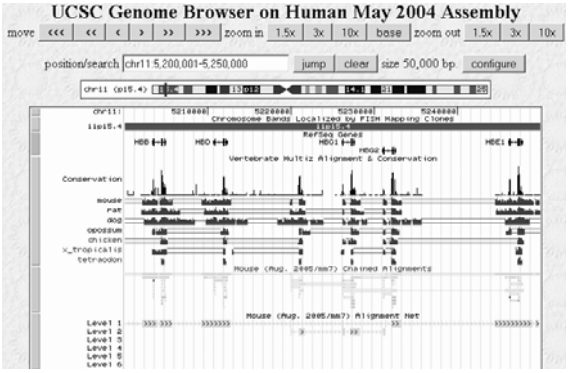
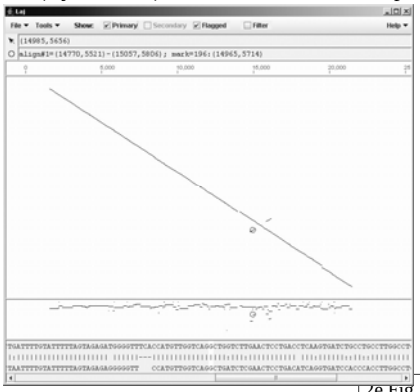


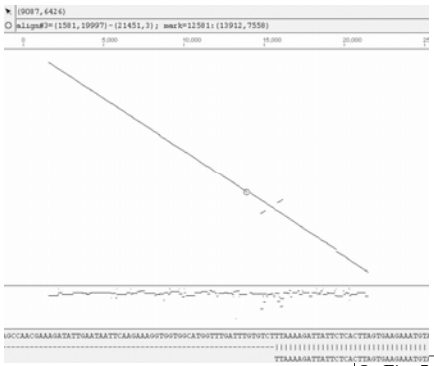
Fig. 5.17

Blastz (laj software): human versus rhesus duplication



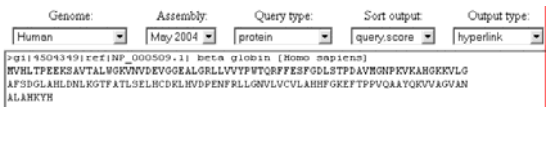
2e Fig. 5.18

Blastz (laj software): human versus rhesus gap



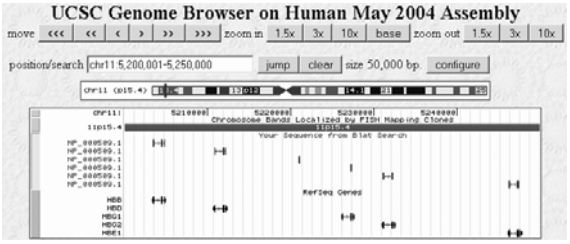
2e Fig. 5.18

BLAT



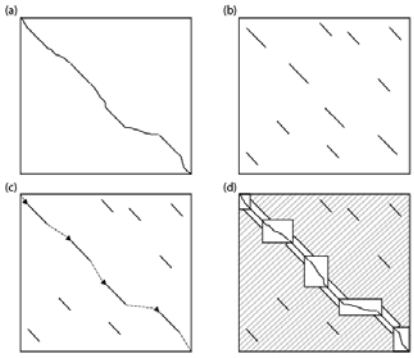
2e Fig. 5.20

BLAT



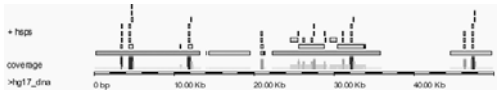
2e Fig. 5.20

LAGAN



2e Fig. 5.21

SSAHA



Outline of today's lecture

Specialized BLAST sites

Finding distantly related proteins: PSI-BLAST
PHI-BLAST

Profile Searches: Hidden Markov models

BLAST-like tools for genomic DNA
PatternHunter
Megablast
BLAT
BLASTZ

BLAST for gene discovery: Find-a-gene

BLAST for gene discovery

You can use BLAST to find a "novel" gene

BLAST for gene discovery

You can use BLAST to find a "novel" gene

You will need to do this for 40% of your grade.

In the first 8 years of this course,
everyone has succeeded at this exercise.

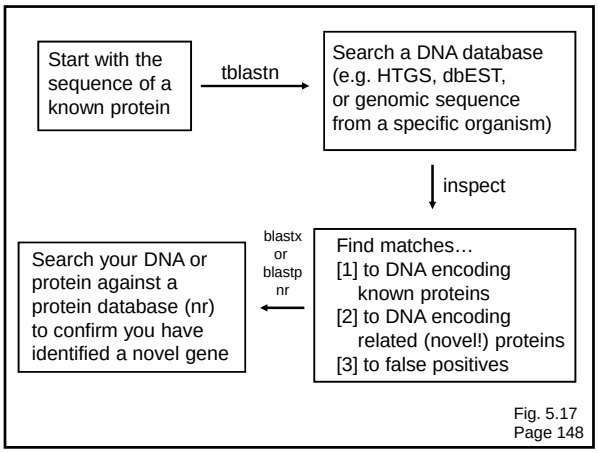


Fig. 5.17
Page 148

Find-a-gene project: HBB *Aquilegia*

2e Fig. 5.23a

Find-a-gene project: *Aquilegia_Arabi*

2e Fig. 5.23b

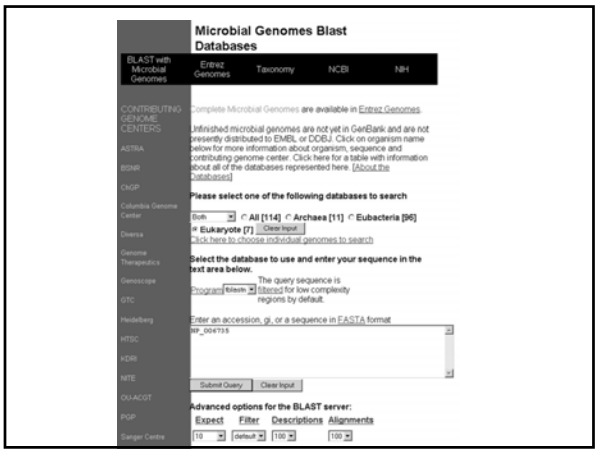
Find-a-gene project: HBB *Vitis*

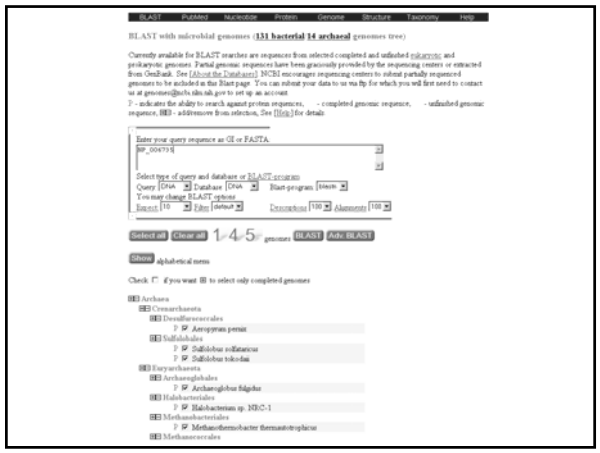
Find-a-gene project: myoglobin tblastn

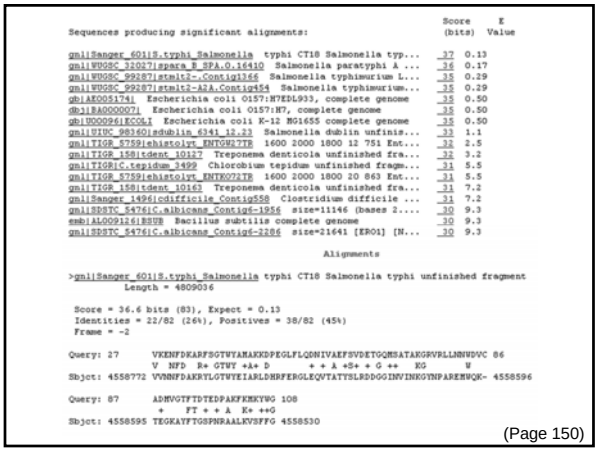
2e Fig. 5.24a

Find-a-gene project: rice myoglobin blastp

2e Fig. 5.24b

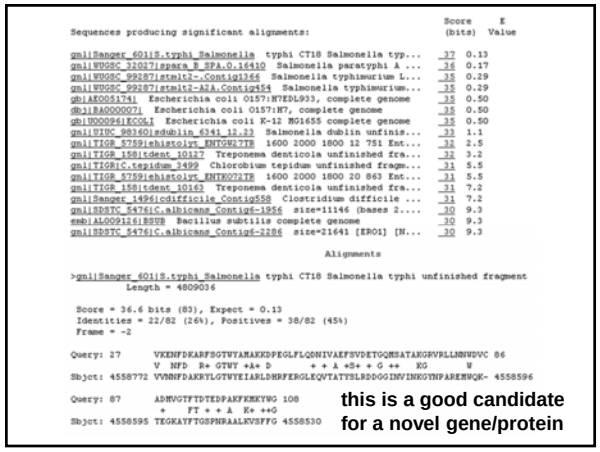






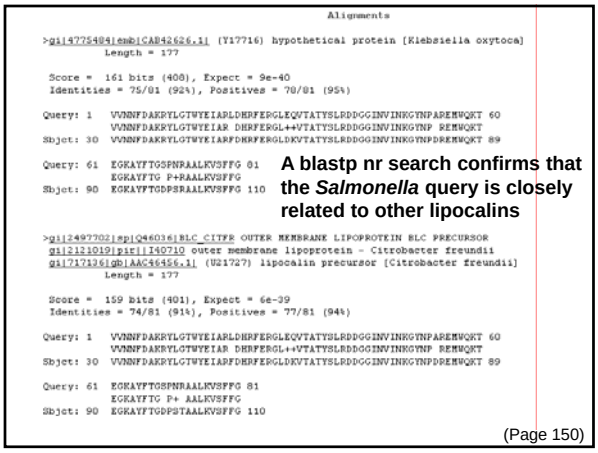
```
                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

gnl|Sanger_601|S.typhi_Salmonella  typhi CT18 Salmonella typ...    37  0.13
gnl|WUGSC_32027|spara_B_SPA.0.16410  Salmonella paratyphi A ...    36  0.17
gnl|WUGSC_99287|stmit2-.Contig1366  Salmonella typhimurium L...    35  0.29
gnl|WUGSC_99287|stmit2-A2A.Contig454  Salmonella typhimurium...    35  0.29
gb|AE005174|  Escherichia coli O157:H7EDL933, complete genome      35  0.50
db|BA000007|  Escherichia coli O157:H7, complete genome            35  0.50
gb|U00096|ECOLI  Escherichia coli K-12 MG1655 complete genome      35  0.50
gnl|UTHC_98360|sdublin_6341_12.23  Salmonella dublin unfinis...    33  1.1
gnl|TIGR_5759|shistolyt_ENTGW27TB  1600 2000 1800 12 751 Ent...    32  2.5
gnl|TIGR_158|tdent_10127  Treponema denticola unfinished fra...    32  3.2
gnl|TIGRIC.tepidum_3499  Chlorobium tepidum unfinished fragme...   31  5.5
gnl|TIGR_5759|shistolyt_ENTKO72TB  1600 2000 1800 20 863 Ent...   31  5.5
gnl|TIGR_158|tdent_10163  Treponema denticola unfinished fra...    31  7.2
gnl|Sanger_1496|cdifficile_Contig558  Clostridium difficile ...    31  7.2
gnl|SDSTC_5476|C.albicans_Contig6-1956  size=11146 (bases 2....    30  9.3
emb|AL009126|BSUB  Bacillus subtilis complete genome               30  9.3
gnl|SDSTC_5476|C.albicans_Contig6-2286  size=21641 [EBO1] [N...    30  9.3

                                Alignments

>gnl|Sanger_601|S.typhi_Salmonella typhi CT18 Salmonella typhi unfinished fragment
          Length = 4809036

 Score = 36.6 bits (83), Expect = 0.13
 Identities = 22/82 (26%), Positives = 38/82 (45%)
 Frame = -2

Query: 27      VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
                V  NFD  R+ GTWY +A+ D       + + A +S+ + G ++   KG       W
Sbjct: 4558772 VVNNFDAKRYLGTWYEIARLDHRFERGLEQVTATYSLRDDGGINVINKGYNPAREMWQK- 4558596

Query: 87      ADMVGTFTDTEDPAKFKMKYWG 108
                 +    FT + + A  K+ ++G
Sbjct: 4558595 TEGKAYFTGSPNRAALKVSFFG 4558530
```

(Page 150)



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

gnl|Sanger_601|S.typhi_Salmonella  typhi CT18 Salmonella typ...    37  0.13
gnl|WUGSC_32027|spara_B_SPA.0.16410  Salmonella paratyphi A ...    36  0.17
gnl|WUGSC_99287|stmit2-.Contig1366  Salmonella typhimurium L...    35  0.29
gnl|WUGSC_99287|stmit2-A2A.Contig454  Salmonella typhimurium...    35  0.29
gb|AE005174|  Escherichia coli O157:H7EDL933, complete genome      35  0.50
db|BA000007|  Escherichia coli O157:H7, complete genome            35  0.50
gb|U00096|ECOLI  Escherichia coli K-12 MG1655 complete genome      35  0.50
gnl|UTHC_98360|sdublin_6341_12.23  Salmonella dublin unfinis...    33  1.1
gnl|TIGR_5759|shistolyt_ENTGW27TB  1600 2000 1800 12 751 Ent...    32  2.5
gnl|TIGR_158|tdent_10127  Treponema denticola unfinished fra...    32  3.2
gnl|TIGRIC.tepidum_3499  Chlorobium tepidum unfinished fragme...   31  5.5
gnl|TIGR_5759|shistolyt_ENTKO72TB  1600 2000 1800 20 863 Ent...   31  5.5
gnl|TIGR_158|tdent_10163  Treponema denticola unfinished fra...    31  7.2
gnl|Sanger_1496|cdifficile_Contig558  Clostridium difficile ...    31  7.2
gnl|SDSTC_5476|C.albicans_Contig6-1956  size=11146 (bases 2....    30  9.3
emb|AL009126|BSUB  Bacillus subtilis complete genome               30  9.3
gnl|SDSTC_5476|C.albicans_Contig6-2286  size=21641 [EBO1] [N...    30  9.3

                                Alignments

>gnl|Sanger_601|S.typhi_Salmonella typhi CT18 Salmonella typhi unfinished fragment
          Length = 4809036

 Score = 36.6 bits (83), Expect = 0.13
 Identities = 22/82 (26%), Positives = 38/82 (45%)
 Frame = -2

Query: 27      VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
                V  NFD  R+ GTWY +A+ D       + + A +S+ + G ++   KG       W
Sbjct: 4558772 VVNNFDAKRYLGTWYEIARLDHRFERGLEQVTATYSLRDDGGINVINKGYNPAREMWQK- 4558596

Query: 87      ADMVGTFTDTEDPAKFKMKYWG 108
                 +    FT + + A  K+ ++G
Sbjct: 4558595 TEGKAYFTGSPNRAALKVSFFG 4558530
```

this is a good candidate for a novel gene/protein



```
                                Alignments

>gi|4977548|emb|CAB42626.1| (Y17716) hypothetical protein [Klebsiella oxytoca]
          Length = 177

 Score =  161 bits (408), Expect = 9e-40
 Identities = 75/81 (92%), Positives = 78/81 (95%)

Query: 1   VVNNFDAKRYLGTWYEIARLDHRFERGLEQVTATYSLRDDGGINVINKGYNPAREMWQKT 60
           VVNNFDAKRYLGTWYEIAR DHRFERGL++VTATYSLRDDGGINVINKGYNP REMWQKT
Sbjct: 30  VVNNFDAKRYLGTWYEIARFDHRFERGLDKVTATYSLRDDGGINVINKGYNPDREMWQKT 89

Query: 61  EGKAYFTGSPNRAALKVSFFG 81
           EGKAYFTG P+RAALKVSFFG
Sbjct: 90  EGKAYFTGDPSRAALKVSFFG 110

>gi|24977021|sp|Q460361BLC_CITFR OUTER MEMBRANE LIPOPROTEIN BLC PRECURSOR
 gi|2121019|pir||I40710 outer membrane lipoprotein - Citrobacter freundii
 gi|1717126|gb|AAC46456.1| (U21727) lipocalin precursor [Citrobacter freundii]
          Length = 177

 Score =  159 bits (401), Expect = 6e-39
 Identities = 74/81 (91%), Positives = 77/81 (94%)

Query: 1   VVNNFDAKRYLGTWYEIARLDHRFERGLEQVTATYSLRDDGGINVINKGYNPAREMWQKT 60
           VVNNFDAKRYLGTWYEIAR DHRFERGL++VTATYSLRDDGGINVINKGYNP REMWQKT
Sbjct: 30  VVNNFDAKRYLGTWYEIARFDHRFERGLDKVTATYSLRDDGGINVINKGYNPDREMWQKT 89

Query: 61  EGKAYFTGSPNRAALKVSFFG 81
           EGKAYFTG P+ AALKVSFFG
Sbjct: 90  EGKAYFTGDPSTAALKVSFFG 110
```

A blastp nr search confirms that the *Salmonella* query is closely related to other lipocalins

(Page 150)

BLAST for gene discovery

You can use BLAST to find a “novel” gene

Ideally, try to find a new gene this week or next.
I will provide sample projects from last year.