

Advanced Database Searching

February 13, 2008

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

Copyright notice

Many of the images in this powerpoint presentation are from *Bioinformatics and Functional Genomics* by Jonathan Pevsner (ISBN 0-471-21004-8). Copyright © 2003 by John Wiley & Sons, Inc.

These images and materials may not be used without permission from the publisher. We welcome instructors to use these powerpoints for educational purposes, but please acknowledge the source.

The book has a homepage at <http://www.bioinfbook.org> including hyperlinks to the book chapters.

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

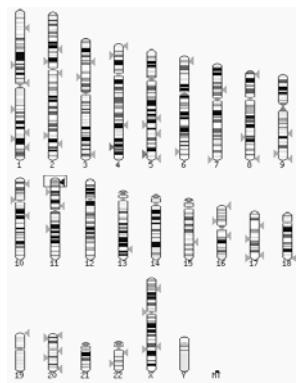
Molecule-specific BLAST sites

Specialized algorithms (WU-BLAST 2.0)

Page 135

Fig. 5.1
Page 131

Ensembl BLAST output includes an ideogram



2e Fig. 5.1

Links	Query Start	End	Ori	Chromosome Name	Start	End	Ori	Chromosome Name	Start	End	Ori	Stats	Score	E-val	P-val	%D	Length
[AITS EG5]	31	100	*	Chr:11	5211000	5204007	*	Chr:11	5211000	5204007	*	602	1.2e-100	4.2e-100	76	76	100
[AITS EG5]	31	124	*	Chr:11	5211721	5212621	*	Chr:11	5211721	5212621	*	602	1.2e-100	4.2e-100	81	81	100
[AITS EG5]	31	124	*	Chr:11	5232000	5232322	*	Chr:11	5232000	5232322	*	532	7.4e-89	7.4e-89	75	75	81
[AITS EG5]	13	121	*	Chr:11	5247102	5247566	*	Chr:11	5247102	5247566	*	529	2.4e-43	3.4e-43	56	56	820
[AITS EG5]	13	121	*	Chr:11	5247102	5247566	*	Chr:11	5247102	5247566	*	529	2.4e-43	3.4e-43	56	56	820
[AITS EG5]	32	104	*	Chr:11	5220915	5221133	*	Chr:11	5220915	5221133	*	436	1.2e-79	1.2e-79	72	72	81
[AITS EG5]	101	101	*	Chr:11	5263400	5263538	*	Chr:11	5263400	5263538	*	368	1.2e-100	1.2e-100	91	91	87
[AITS EG5]	147	147	*	Chr:11	5264673	5264904	*	Chr:11	5264673	5264904	*	323	1.2e-89	1.2e-89	84	84	84
[AITS EG5]	1	45	*	Chr:11	5264699	5264827	*	Chr:11	5264699	5264827	*	272	1.8e-45	1.8e-45	88	88	85
[AITS EG5]	109	147	*	Chr:11	5264828	5264928	*	Chr:11	5264828	5264928	*	263	2.5e-29	2.5e-29	74	74	83
[AITS EG5]	147	147	*	Chr:11	5264928	5264934	*	Chr:11	5264928	5264934	*	263	2.5e-29	2.5e-29	74	74	83
[AITS EG5]	31	143	*	Chr:16	166926	167237	*	Chr:16	166926	167237	*	280	1.2e-19	1.2e-19	35	35	54
[AITS EG5]	31	143	*	Chr:16	167122	167433	*	Chr:16	167122	167433	*	256	3.1e-19	3.1e-19	35	35	118

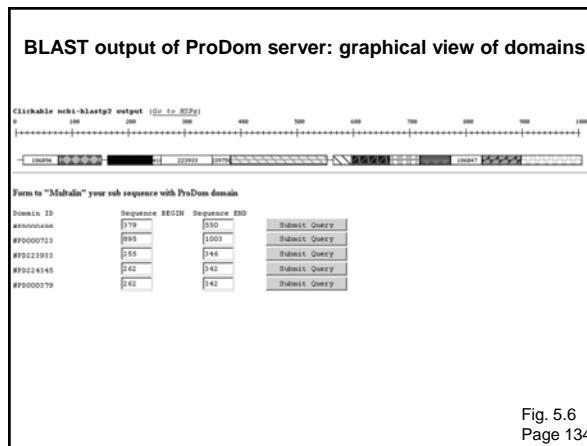
2e Fig. 5.2

General DNA and Protein Searches	
Tool	Description
Blast2-WU Protein	Washington University (WU) blast2 for protein databases. (blast 2.0 with gaps)
Blast2-WU Nucleotide	Washington University (WU) blast2 for nucleotide databases. (blast 2.0 with gaps)
Blast2-HCBI Protein	NCBI blast2 (blastall) program for protein databases.
Blast2-HCBI Nucleotide	NCBI blast2 (blastall) program for nucleotide databases.
Blast2-HCBI EVL	European blast2 Vector Searches. Check your sequences for vector contamination.
PSI-Blast	Position specific Iterative Blast (PSI-Blast) refers to a feature of Blast 2.0 in which a profile is automatically constructed from the first set of Blast alignments.
PHI-Blast	Pattern Hit Initiated Blast (PHI-Blast) treats two occurrence of the same pattern in the query sequence as two independent sequences.
Fasta Nucleotide	Sequence similarity searching against nucleotide databases using Fasta.
Fasta Protein	Sequence similarity searching against protein databases using Fasta.
Fasta Proteome server	Completed Proteomes Fasta server.
Fasta-Genome server	Completed Genomes Fasta server.
Fasta-WGS server	Whole genome shotgun (WGS) Fasta server.

Rigorous Protein Searches	
Tool	Description
MPSearch	AwazuBio, formerly Edinburgh Biocomputing Systems' very fast implementation of the true Smith and Waterman algorithm.
Scansp2.3	NEW Version 2.3 of Scansp2.3 implementation of the Smith-Waterman algorithm for protein database searches.

2e Fig. 5.3

TIGR BLAST

Fig. 5.3
Page 132Fig. 5.6
Page 134

EMBL European Bioinformatics Institute

WU-Blast2 [Help Tools ERN Home RUN BLAST RESET FORM]

YOUR EMAIL: SEARCH TITLE: RESULTS DATABASE: PROGRAM:

MATRIX: DNA STRAND: EXP. THR: FILTER: VIEW FILTER

HISTOGRAM: STATS: SORT: SCORES: ALIGNMENTS:

Enter or Paste a PROTEIN Sequence in any format:
Upload a file: Browse...
RUN BLAST RESET FORM

This document was last modified on: Thursday, July 05, 2001 10:52:03
Comments or suggestions report@ebi.ac.uk
© EBI 2000

If you plan to use these services during a course please contact us using the email above.

Fig. 5.7
Page 136

PROGRAM: Blast (DNA Query vs. DNA DB)
DATABASE: DNA DB

SEARCH: DNA DB

Analysis: CLUSTALW, LIBRARY, Others

Division: ON OFF

QUERY SEQUENCE: Enter sequence
File Upload: Browse...

COPY & PASTE: or

RESULT:

Fig. 5.8
Page 137

Search with FASTA

I am coming from pc190-73.kenneckyriege.org/. I am going to fasta.bisch.virginia.edu/

Choose program and database(s) to query:

Program: FASTA	FASTA
Databases: FASTX	Genbank DNA
<input type="checkbox"/> protein FASTX	<input type="checkbox"/> Primate
<input type="checkbox"/> protein FASTS	<input type="checkbox"/> MTO (High throughput)
<input type="checkbox"/> protein FASTG	<input type="checkbox"/> Rodent
<input type="checkbox"/> protein FASTA	<input type="checkbox"/> Synthetic
<input type="checkbox"/> protein FASTT	<input type="checkbox"/> Viral
<input type="checkbox"/> protein FASTS (mg)	<input type="checkbox"/> Other Mammals
<input type="checkbox"/> protein FASTG (mg)	<input type="checkbox"/> Vertebrates
<input type="checkbox"/> NCBI/blast NR	<input type="checkbox"/> TIGR Human gene index (validated)
<input type="checkbox"/> NCBI/blast NR (reg)	<input type="checkbox"/> TIGR Mouse gene index (validated)
<input type="checkbox"/> NCBI/blast PDB structures	<input type="checkbox"/> Invertebrates
<input type="checkbox"/> NCBI/blast PDB sequences	<input type="checkbox"/> Plants
<input type="checkbox"/> NCBI/blast PDB (reg)	<input type="checkbox"/> BLAST human ESTs
<input type="checkbox"/> NCBI/blast Swiss-Prot (reg)	<input type="checkbox"/> Bacteria
<input type="checkbox"/> NCBI/blast Swiss-Prot	<input type="checkbox"/> BLAST mouse ESTs
<input type="checkbox"/> NCBI/blast WORM/PEP	<input type="checkbox"/> Phage
<input type="checkbox"/> Drosophila proteome	
<input type="checkbox"/> Yeast NR	

Search individual Proteomes/Genomes

Protein DNA (both-strands) DNA (forward only) DNA (rev-comp only)

Enter query sequence: **FASTA format** Subset range:

Enter protein sequence browser

Enter DNA sequence browser

Submit Query

Other options:

King: Protein matrix: DNA matrix: gap: ext: misc:
 Default Default

FASTA server at the University of Virginia

NCBI Conserved Domain Database - Webpage

Help | Home | Search | Communicate | Help | Log in | Log out | Help | Search | Communicate | Help | Log in | Log out

CDD Version: 20.07 (2007-08-01) Last updated: 2007-08-01

K CDD: A Conserved Domain Database and Search Service

Conserved domains often contain several sub-domains, each with a distinct evolutionary origin and function. The CDD search service may be used to search for a specific domain or for all domains contained in a sequence.

To identify conserved domains in a protein sequence, the CDD search service applies the reverse position-specific BLAST algorithm. The query sequence is compared against a database of known domains. Domains containing conserved domain alignments will be displayed as a pairwise alignment between the query sequence with a representative domain sequence, or as a multiple alignment.

Run CDD search

Search Database: **20.07** Search protein or PDB ID: or ID: Submit Query

Search CDD: Search CDD

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

730496 66 FTVDENGQMSATAKGRVRLFNNWVCAFDIGSFTDTEPAFKHMYNGVASFLQRGNDDH 125
 200679 63 FSVDEKGHMSATAKGRVRLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 122
 206589 34 FSVDEKGHMSATAKGRVRLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 93
 2136812 2 MSATAKGRVRLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 53
 132408 65 FKIENGDGTTATAKGRVRLDKELECLANHV/GTFITNDPDKFPMYHVGALALILERGLDDH 124
 267584 44 FSVDESGNVTATAHGRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 103
 267585 44 FSVDESGNVTATAHGRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 103
 8777608 63 FTIHEDGAMTTATAKGRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 122
 6687453 60 FKVDEDGTRATA1GRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 119
 10697027 81 FKVQDGDTATATGRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 140
 13645517 1 RVGTFDTEDPAFKHMYNGVASFLQRGNDDH 32
 13925316 38 FSVDGSQKMTATAQGRVILLLNNEVCAFDIVGTFDTEDPAFKHMYNGVASFLQRGNDDH 97
 131649 65 YVVEDGTMASSKGRVQLFGVVICADHAAQTTDPTTPAKHMYTYQGLASLYSGDNY 126

↑ R,I,K ↑ C ↑ D,E,T ↑ K,R,T ↑ N,L,Y,G

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	-2	-3	-2	1	-4	-3	-3	12	2	2	-3
4 V	0	-3	-3	-4	-1	-3	-3	-4	-3	1	-3	1	-1	-3	-2	0	-3	-1	4	
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	2	-3
6 A	5	-2	-2	-2	-2	-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	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	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12 A	5	-2	-2	-2	-2	-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	0	-2	-3	1	3	0	-3	-2	-2	
16 A	4	-2	-1	-2	-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	-2	-1	-1	-1	-2	-1	1	5	-3	-2	0			
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	2	-3
41 Y	-2	-2	-2	-3	-3	-2	-2	3	-2	-1	-2	-1	3	-3	-2	-2	7	2	7	-1
42 A	4	-2	-2	-2	-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	-3	-2	1	2	-2	6	0	-3	-2
2 K	-1	1	0	1	-4	2	4	-2	0	-3	3	-2	-4	-1	0	-1	-3	-2	-3	
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3		
4 V	0	-3	-3	-4	-5	-3	-2	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3	
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3		
6 A	5	-2	-2	-2	-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	
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	2	0	
9 L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	
10 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	
11 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	
12 A	5	-2	-2	-2	-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	
14 A	3	-2	-1	-2	-1	-1	2	-2	-2	-1	-2	-1	-3	-1	1	-1	-3	-3	-1	
15 A	2	-1	0	-1	-2	2	0	2	-1	-3	0	-2	-3	1	3	0	-3	-2	-2	
16 A	4	-2	-1	-2	-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	-2	-1	-1	-2	-1	-1	-2	-1	1	5	-3	-2	0	
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-2	-3	-2	1	-4	-3	12	2	-3		
41 Y	-2	-2	-2	-3	-3	-2	-2	3	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1	
42 A	4	-2	-2	-2	-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

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	> 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

Sequences with E-value BETTER than threshold

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	398	1e-111	UG
refINP_001630.11	apolipoprotein D precursor [Homo sapiens]	57.4	7e-09	UG
refINP_001018059.11	glycodelin precursor [Homo sapiens] >ref!...	26.2	0.019	UG
refINP_001624.11	alpha-1-microglobulin/bikunin precursor [Homo s...	35.8	0.021	UG

Sequences producing significant alignments:

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	398	1e-111	UG
refINP_001630.11	apolipoprotein D precursor [Homo sapiens]	57.4	7e-09	UG
refINP_001018059.11	glycodelin precursor [Homo sapiens] >ref!...	26.2	0.019	UG
refINP_001624.11	alpha-1-microglobulin/bikunin precursor [Homo s...	35.8	0.021	UG

Run PSI-BLAST iteration 2

Sequences with E-value WORSE than threshold

	Score	E	(Bits)	Value
refINP_000597.11	complement component C9, gamma polypeptide [Homo sapiens]	33.9	0.077	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	28.5	3.8	UG
refINP_066015.21	hypothetical protein LOC57724 [Homo sapiens]	27.3	7.5	UG

Run PSI-BLAST iteration 2

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

Sequences with E-value BETTER than threshold

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	368	1e-102	UG
refINP_001630.11	apolipoprotein D precursor [Homo sapiens]	149	2e-36	UG
refINP_001018059.11	glycodelin precursor [Homo sapiens] >ref!...	134	5e-32	UG
refINP_001624.11	alpha-1-microglobulin/bikunin precursor [Homo s...	125	2e-29	UG
refINP_0011159497.11	PREDICTED: similar to Glycodelin precursor...	70.6	1e-12	UG
refINP_000945.21	PREDICTED: similar to Glycodelin precursor (...	69.3	2e-12	UG
refINP_000945.21	prostaglandin H2 D-isomerase [Homo sapiens]	43.3	1e-04	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	39.6	0.002	UG
refINP_045564.21	lipocalin 9 [Homo sapiens]	28.8	0.002	UG
refINP_001018059.11	lipocalin 9 [Homo sapiens]	28.5	0.003	UG
refINP_005597.11	complement component C9, gamma polypeptide [Homo sapiens]	26.9	0.010	UG

Sequences producing significant alignments:

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	368	1e-102	UG
refINP_001630.11	apolipoprotein D precursor [Homo sapiens]	149	2e-36	UG
refINP_001018059.11	glycodelin precursor [Homo sapiens] >ref!...	134	5e-32	UG
refINP_001624.11	alpha-1-microglobulin/bikunin precursor [Homo s...	125	2e-29	UG
refINP_0011159497.11	PREDICTED: similar to Glycodelin precursor...	70.6	1e-12	UG
refINP_000945.21	PREDICTED: similar to Glycodelin precursor (...	69.3	2e-12	UG
refINP_000945.21	prostaglandin H2 D-isomerase [Homo sapiens]	43.3	1e-04	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	39.6	0.002	UG
refINP_045564.21	lipocalin 9 [Homo sapiens]	28.8	0.002	UG
refINP_001018059.11	lipocalin 9 [Homo sapiens]	28.5	0.003	UG
refINP_005597.11	complement component C9, gamma polypeptide [Homo sapiens]	26.9	0.010	UG

Run PSI-BLAST iteration 3

Sequences with E-value WORSE than threshold

	Score	E	(Bits)	Value
refINP_0022588.11	lipocalin 1 precursor [Homo sapiens]	31.5	0.48	UG
refINP_004534.21	nebulin [Homo sapiens]	30.4	1.0	UG
refINP_775903.21	zinc finger protein 776 [Homo sapiens]	30.0	1.2	UG
refINP_055993.11	hypothetical protein LOC23211 [Homo sapiens]	29.2	2.1	C
refINP_001018059.11	diaphanous homolog 3 isoform a [Homo sapiens]	28.8	2.5	UG
refINP_060146.21	zinc finger, HEZ domain containing [Homo sapi...	28.4	3.4	UG
refINP_055997.11	zinc finger, HEZ domain containing protein 2A precursor [Homo sapiens]	28.1	4.4	UG
refINP_945134.11	odorant binding protein 2A precursor [Homo sapi...	27.3	9.3	UG

Run PSI-BLAST iteration 3

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

Sequences with E-value BETTER than threshold

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	358	2e-99	UG
refINP_000597.11	complement component C9, gamma polypeptide [Homo sapiens]	33.9	0.077	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	28.5	3.8	UG
refINP_066015.21	hypothetical protein LOC57724 [Homo sapiens]	27.3	7.5	UG

Sequences producing significant alignments:

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	358	2e-99	UG
refINP_000597.11	complement component C9, gamma polypeptide [Homo sapiens]	33.9	0.077	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	28.5	3.8	UG
refINP_066015.21	hypothetical protein LOC57724 [Homo sapiens]	27.3	7.5	UG

Run PSI-BLAST iteration 4

Sequences with E-value WORSE than threshold

	Score	E	(Bits)	Value
refINP_000597.11	PREDICTED: similar to Glycodelin precursor...	60.8	8e-10	UG
refINP_044662.21	PREDICTED: similar to Glycodelin precursor (...	60.8	8e-10	UG
refINP_000945.21	PREDICTED: similar to Glycodelin precursor (...	60.8	8e-10	UG
refINP_000945.21	prostaglandin H2 D-isomerase [Homo sapiens]	59.1	4e-09	UG
refINP_045564.21	lipocalin 8 [Homo sapiens]	41.7	2e-04	UG
refINP_045564.21	lipocalin 8 [Homo sapiens]	41.7	2e-04	UG
refINP_045564.21	lipocalin 6 [Homo sapiens]	41.7	2e-04	UG
refINP_045564.21	lipocalin 6 [Homo sapiens]	41.7	2e-04	UG
refINP_055997.11	odorant binding protein 2A precursor [Homo sapi...	38.5	0.003	UG
refINP_055997.11	odorant binding protein 2B [Homo sapiens]	36.5	0.016	UG
refINP_0022588.11	lipocalin 1 precursor [Homo sapiens]	34.9	0.039	UG

Run PSI-BLAST iteration 4

PSI-BLAST search: human RBP versus RefSeq, iteration 4

Sequences with E-value BETTER than threshold

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	358	2e-99	UG
refINP_000597.11	complement component C9, gamma polypeptide [Homo sapiens]	33.9	0.077	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	28.5	3.8	UG
refINP_066015.21	hypothetical protein LOC57724 [Homo sapiens]	27.3	7.5	UG

Sequences producing significant alignments:

	Score	E	(Bits)	Value
refINP_006735.21	retinol-binding protein 4, plasma precursor [Homo sapiens]	358	2e-99	UG
refINP_000597.11	complement component C9, gamma polypeptide [Homo sapiens]	33.9	0.077	UG
refINP_076222.11	MSFL2541 [Homo sapiens]	28.5	3.8	UG
refINP_066015.21	hypothetical protein LOC57724 [Homo sapiens]	27.3	7.5	UG

Run PSI-BLAST iteration 4

Sequences with E-value WORSE than threshold

	Score	E	(Bits)	Value
refINP_040603.21	lipocalin 12 [Homo sapiens]	31.1	0.66	UG
refINP_00101712.21	lipocalin 10 [Homo sapiens]	30.7	0.82	UG
refINP_00101712.21	lipocalin 10 [Homo sapiens]	30.7	0.82	UG
refINP_040603.21	lipocalin 8 [Homo sapiens]	30.3	0.99	UG
refINP_040603.21	segrin 4 isoform 1 [Homo sapiens]	30.3	0.99	UG
refINP_240455.11	phosphodiesterase 3A isoform 3 [Homo sapiens]	27.6	5.9	UG
refINP_240455.11	phosphodiesterase 3A isoform 1 [Homo sapiens]	27.6	5.9	UG
refINP_240455.11	phosphodiesterase 3A isoform 2 [Homo sapiens]	27.6	5.9	UG
refINP_030174.21	gamma-butyryl ester diacylglycerase [Homo sapiens]	27.2	8.5	UG
refINP_030174.21	nebulin [Homo sapiens]	27.2	9.6	UG

Run PSI-BLAST iteration 4

RBP4 match to ApoD, PSI-BLAST iteration 2

>refINP_001630.11 **UG** apolipoprotein D precursor [Homo sapiens]
Length=189

Score = 175 bits (443), Expect = 1e-42, Method: Composition-based stats.
Identities = 45/163 (27%), Positives = 77/163 (47%), Gaps = 31/163 (19%)

Query 14 GSGAAEFDGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Subject 14 AAKGAAEFLGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Length=189

Query 14 GSGAAEFDGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Subject 14 AAKGAAEFLGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Length=189

Query 74 TAX+G V + T + +PAK ++K+ V + S + +W+ T
Subject 74 LNLQLRLADGTWQNQE-----KATPVWLTKEALKLEVYKFSPWMP3-----APWVILATVYNTV
Length=189

Query 120 DYTVAYQVYSCR-----LLNLDGTACDSY3FVFSRDP9NGLPFEA 106
DY+ YA+ YSC L +D+ ++++ +R+P N LPFE
Subject 124 D YVYVSLVYVSCTIIQLFNVD-----FAUTLAPMNP-LPFEET 159

2e Fig. 5.6

RBP4 match to ApoD, PSI-BLAST iteration 3

>refINP_001630.11 **UG** apolipoprotein D precursor [Homo sapiens]
Length=189

Score = 146 bits (368), Expect = 6e-34, Method: Composition-based stats.
Identities = 41/163 (25%), Positives = 76/163 (46%), Gaps = 20/163 (12%)

Query 14 GSGAAEFDGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Subject 14 AAKGAAEFLGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Length=189

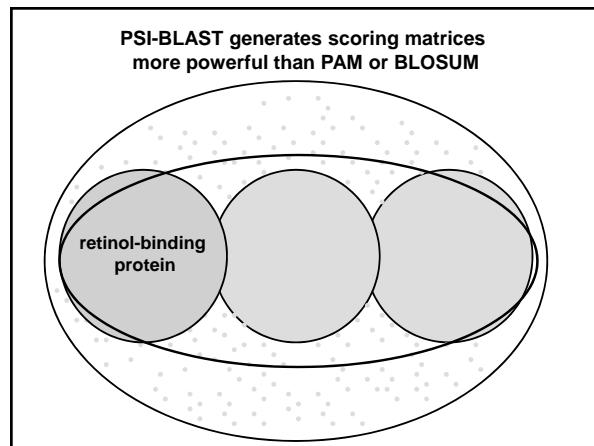
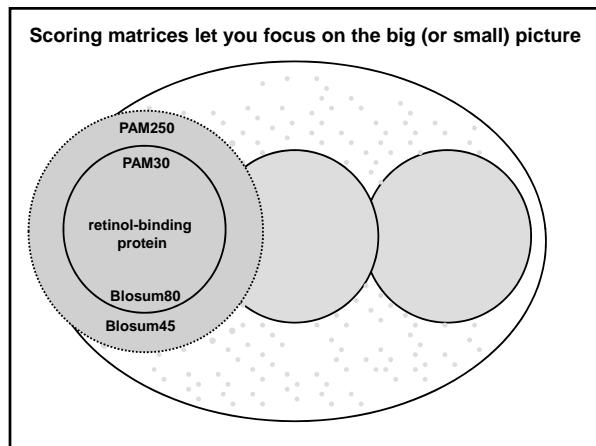
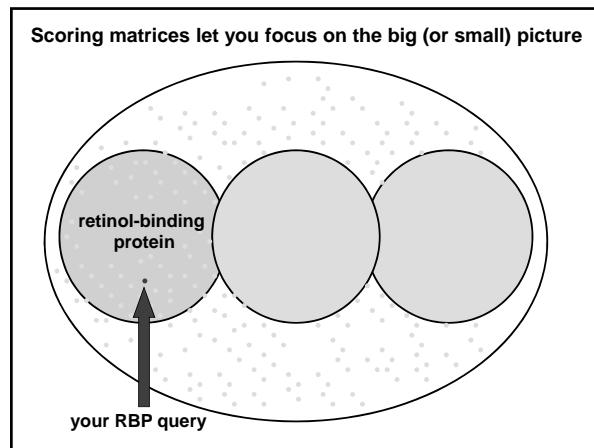
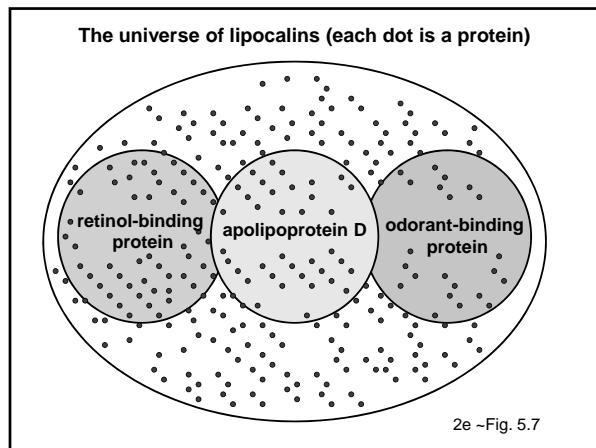
Query 14 GSGAAEFDGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Subject 14 AAKGAAEFLGRVSSPFVWTFKAFRFSVQVAMAMAAV...
Length=189

Query 74 TAX+G V + T + +PAK ++K+ V + S + +W+ T+Y
Subject 74 LNLQLRLADGTWQNQE-----KATPVWLTKEALKLEVYKFSPWMP3-----APWVILATVYNTV
Length=189

Query 120 AYQYSCR-----LLNLDGTACDSY3FVFSRDP9NGLPFEA 171
A+ YSC L +D+ ++++ +R+P N P +
Subject 129 ALVYFSCTIIQLFNVD-----FAUTLAPMNP-LPFEET 165

2e Fig. 5.6

5



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:

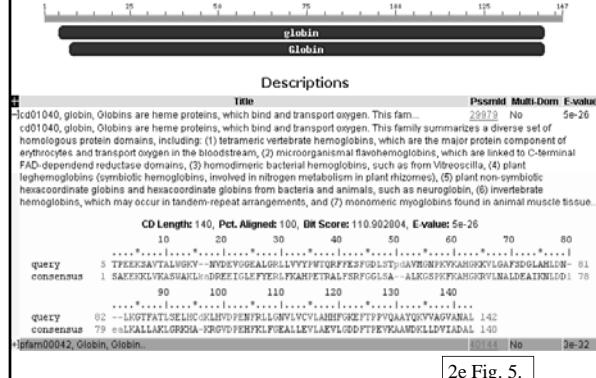
- [1] Apply filtering of biased composition regions
- [2] Adjust E value from 0.001 (default) to a lower value such as E = 0.0001.
- [3] Visually inspect the output from each iteration.
Remove suspicious hits by unchecking the box.

Page 144

<input checked="" type="checkbox"/> g116978531(p)P04009_11 apolipoprotein E [lambda surveyed]...	137	6e-25
<input checked="" type="checkbox"/> g115424973(d)BAA83343e_11 (907752) alpha-microglobulin/human...	155	6e-14
<input checked="" type="checkbox"/> g11539331(p)BAA83220_11 apolipoprotein B, apob [human, plasma, ...]	153	6e-14
<input checked="" type="checkbox"/> g115419692(p)em1CXB4469_11 (302624) RBP (aa 101-172) [Homo sapiens]	139	1e-12
<input checked="" type="checkbox"/> g1154021151(p)IP01630_11 apolipoprotein D precursor [Homo sapi...	138	4e-32
<input checked="" type="checkbox"/> g115047631(p)P171511ARD_RABIT APOLIPOPROTEIN D PRECURSOR g114802...	134	4e-31
<input checked="" type="checkbox"/> g1170232411(p)P190911ARD_CAVPO APOLIPOPROTEIN D PRECURSOR >g111...	132	7e-31
<input checked="" type="checkbox"/> g1129923041(q)BAA202945_11 (AF023394) mutant retinol binding prot...	90	9e-15
<input checked="" type="checkbox"/> g1132440981(q)BAA832393_11 (904649) apolipoprotein D, apob [C-ter...	77	9e-14
<input checked="" type="checkbox"/> g1129923041(q)BAA202945_11 (AF023394) mutant retinol binding prot...	75	1e-11
<input checked="" type="checkbox"/> g112444159(p)P49251(LIZ2_B99) LAMARILLO PROTEIN PRECURSOR g11...	63	1e-09
<input checked="" type="checkbox"/> g1125074511(p)P005971AMB_B07M AMBP PROTEIN PRECURSOR (CONTAINS...)	63	2e-09
<input checked="" type="checkbox"/> g112497692(p)P007454AMB_H008 AMBP PROTEIN PRECURSOR (CONTAINS...)	63	2e-09
<input checked="" type="checkbox"/> g116690581(p)IP01649_11 alpha 1 microglobulin/bikunin [Ran Ni...	62	2e-09
<input checked="" type="checkbox"/> g1123034451(p)BAA202945_11 (AK004007) putative [Run mucous]	62	3e-09
<input checked="" type="checkbox"/> g1149768971(q)BAA832393_11 (AK004007) putative [Run mucous]	62	3e-09
<input checked="" type="checkbox"/> g1125078861(p)P043461AMB_P10 AMBP PROTEIN PRECURSOR (CONTAINS...)	61	8e-09
<input checked="" type="checkbox"/> g1110502071(p)C11J2C556 alpha-1-microglobulin/inter-alpha-trypsin...	60	1e-06
<input checked="" type="checkbox"/> g1129923041(q)BAA202945_11 (AF023394) alpha-1-microglobulin/bikun...	59	2e-08
<input checked="" type="checkbox"/> g1125074511(p)P005971AMB_B07M AMBP PROTEIN PRECURSOR - pas...	59	2e-06
<input checked="" type="checkbox"/> g1118921(p)BAA202945_11 (329097) precursor codes for two protein...	59	2e-06
<input checked="" type="checkbox"/> g119318121(q)BAA832393_11 (AF176500) neural Lamarillo ...	39	3e-08
<input checked="" type="checkbox"/> g1172946031(q)BAA832397_11 (AK003584) MBL gene product [Sromophili...	38	3e-08
<input checked="" type="checkbox"/> g1117373101(p)P8000071CRAB_HORN_CROATACAVANII_A2 SUBUNIT g110275...	37	8e-08
<input checked="" type="checkbox"/> g1124976951(p)P065559AMB_MEAL AMBP PROTEIN PRECURSOR (CONTAINS...)	37	1e-07
<input checked="" type="checkbox"/> g1110504681(q)BAA202945_11 insecticytomin A - tobacco hornworm g11971...	56	1e-07
<input checked="" type="checkbox"/> g114502071(q)BAA832393_11 alpha-1-microglobulin/bikunin precu...	56	2e-07
<input checked="" type="checkbox"/> g11114440091(q)BAA832393_11 (L41649) galerin [Galleria mellonella]	56	2e-07
<input checked="" type="checkbox"/> g1124976951(p)P065559AMB_MEAL AMBP PROTEIN PRECURSOR (CONTAINS...)	55	1e-07
<input checked="" type="checkbox"/> g111255891(p)BAA202945_11 (D89712) Prolipocalatin D Synthesis Enz...	54	5e-07
<input checked="" type="checkbox"/> g115397171(q)BAA832393_11 retinol-binding protein - est (fragment)	54	8e-07
<input checked="" type="checkbox"/> g112454721(q)BAA832393_11 retinol-binding protein, RBP (N-terminal...	53	1e-06
<input checked="" type="checkbox"/> g112454721(q)BAA832393_11 retinol-binding protein, RBP (N-terminal...	52	3e-06
<input checked="" type="checkbox"/> g1110792951(q)BAA832393_11 retinol-binding protein, RBP (N-terminal...	52	3e-06
<input checked="" type="checkbox"/> g117220091(p)P392811BLC_ECOLI OUTER MEMBRANE LIPOPROTEIN BLC_PPE...	51	9e-06

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.

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.

Page 145

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.

Page 145

Align three lipocalins (RBP and two bacterial lipocalins)

```

      1
ecblc MRLLPLVAAA TAAFLVVACAS SPTPPRGVTV VNNFDAKRYL GTWYEIARP
      vc MRAIPLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWEVARLD
hsrbp ---MKWWALL LLLAAWAAA RDCRVSSFRV KENFDKARFS GTIWYAMAKKD

```

Fig. 5.9

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

```

1   MRLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDKRYL GTWYEIARPD
ecblc  vc  MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWEVARLD
hsrbp  ---MKWWAL LLLAAWAAA RDCRVSSFRV KENFDKARFS GTWYAMAKKD

GTWYEI
K AV
M

```

Fig. 5.9

Create a pattern using the appropriate syntax

```

1   MRLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDKRYL GTWYEIARPD
ecblc  vc  MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWEVARLD
hsrbp  ---MKWWAL LLLAAWAAA 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.

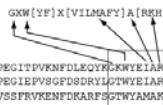
[] means any one of the characters enclosed in the brackets e.g., [LFYT] means one occurrence of L or F or Y or T

- 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

PHI pattern: GXW[YF]X[VILMAFY]A[RKH]

ZP_01613353 -----MKAITTILLITGL-FLLTACTSAPEGITPVKNFDLEQYIKKWEIARPD 47
ZP_01006014 MYLLLENGALAMMAVLRRNFWLVLG-MGLASCTSLPEQIEPVSGEUSDRYLWTWEIARPD 59
human -----MKWWALLLAALGSGRAERDCRVSSFRVKENFTKARSGCTYAMAKKD 48
* : :*.* : . . . : ..** : : *.* :*

2e Fig. 5.9

PHI-BLAST: input



2e Fig. 5.9

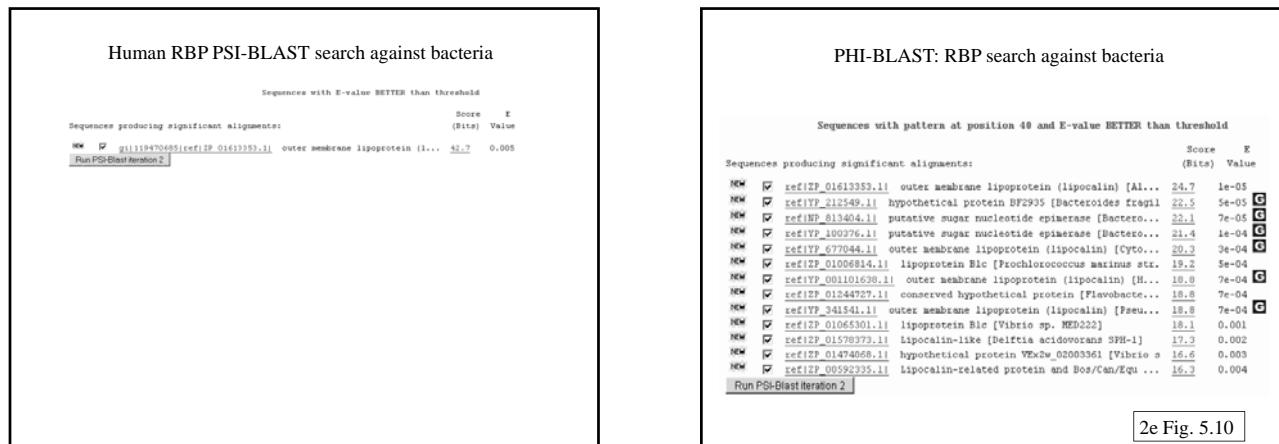
PHI-BLAST: output

```

>ref|ZP_01613353.1| outer membrane lipoprotein (lipocalin) [Alteromonas bacteria
TV-7]
Length=173
Score = 24.7 bits (72), Expect = 1e-05
Identities = 21/90 (26%), Positives = 40/80 (50%), Gaps = 1/90 (1%)
Pattern 31 EHFIAKPSSTEDRRAAGDQPGFLQLQHNTVAEFVSVETGQESTAEGPFLLNNDVYAD 90
Query 91 HVCFTTDEDPKPMQG 110
Sbjct 30 KMFDFEYKGVWEIAPLHSFEFGMEQVTTATTVMDDGIVKVLMGFTTEQKWD-EAE 88
Sbjct 89 GLAKFVEGIDTGHPKV3FFG 108

```

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



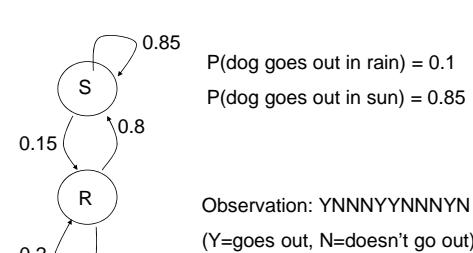
Rain = dog may not want to go outside
Sun = dog will probably go outside

Diagram of a simple Markov model with two states: S (Sun) and R (Rain). Transitions are as follows:
S to S: 0.85
S to R: 0.15
R to S: 0.2
R to R: 0.8

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

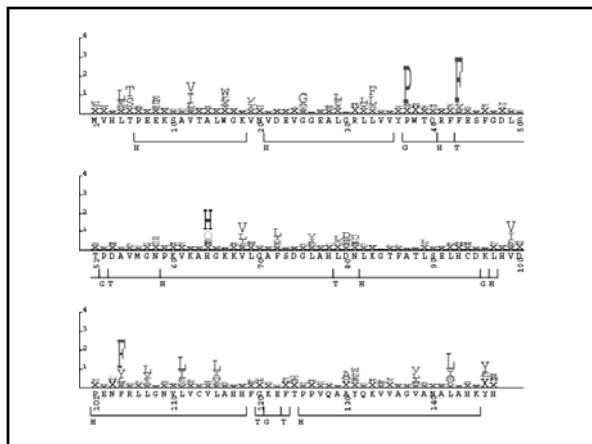
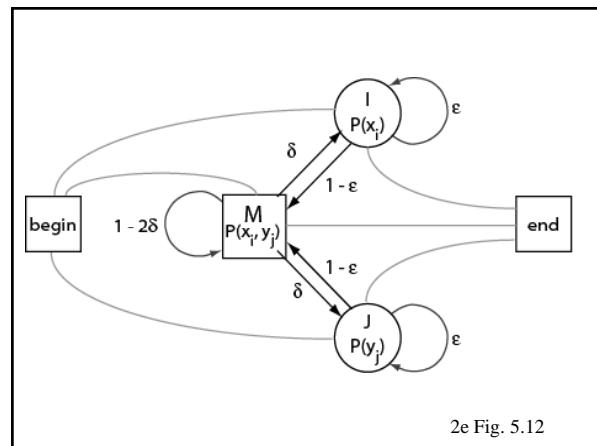
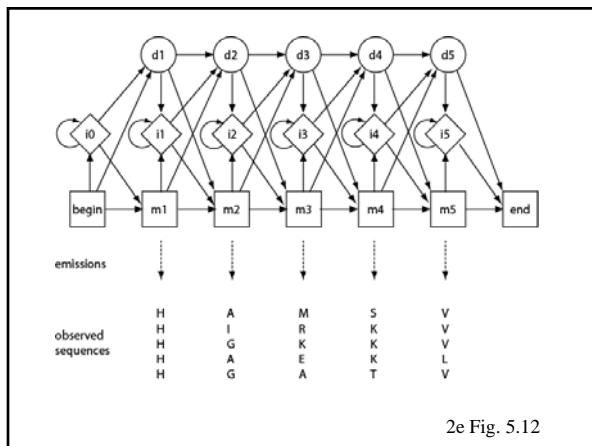
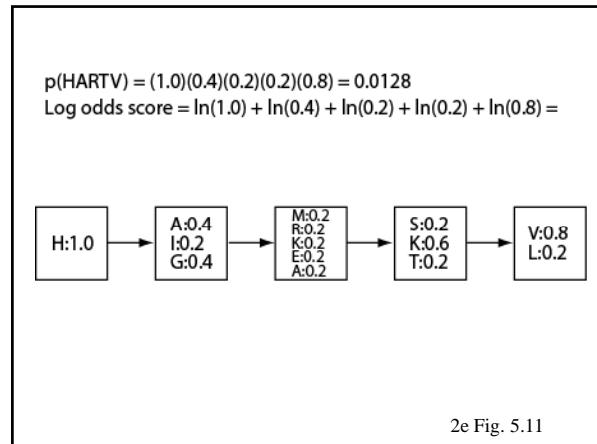
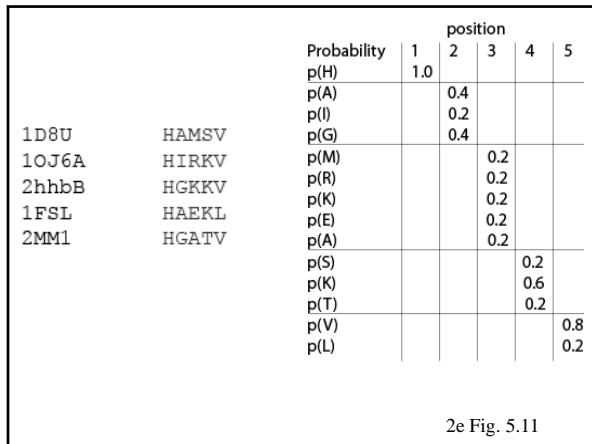
courtesy of Sarah Wheelan

Simple Hidden Markov Model



P(dog goes out in rain) = 0.1
P(dog goes out in sun) = 0.85

Observation: YNNNNYYNNNNYN
(Y=goes out, N=doesn't go out)
What is underlying reality (the hidden state chain)?



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
-*+mkwMkLLLAlaLagvfgaRdrdsvgKcrvpssPRGfrfVknNFDv
mkw+++LLLaA + +aErd Crv+s frVknNFD+
gi|5803139 1 MEKWVALLLALLA-W--AAARD-----CRVSS---FRVKENFDK 33
erylgTwYeIakDkprFGrGLlgdkItAeySlEeHgMsataeGri+rVL
+r+GhY+-+aRkDp E Gl+ldg+i+A+S+E+G+Msata+Gri+rL
gi|5803139 34 ARFGSGTWTYAMAERKDp-E-GLFLQDNIVAEFSVDETQMSATAKGVRVRL 80
eNKeIcAdVkvGtvtqjEGeaseevLtadPaklk1KyavGvaSflqpGfdy
+N++cAD+-VGt+I++E dPak+k+Ky+GvaSflq+Gdd+
gi|5803139 81 NNNDVCAJDMVGTFTDE-----DPAKFKMKTWGVASFLQKGNDNH 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
-*+mkwMkLLLAlaLagvfgaRdrdsvgKcrvpssPRGfrfVknNFDv
N+LL+ +A a ++ Af+v+cC++p+PF+G+v+NFD+
gi|1676758 1 ---MRLLPVVA-----AVTA-AFLUVACSSPFPKGVTVNNNDA 36
erylgTwYeIakDkprFGrGLlgdkItAeySlEeHgMsataeGri+rVL
+r+GhY+-+aRkDp E Gl+ldg+i+A+S+E+G+Msata+Gri+rL
gi|1676758 37 KRYLGWTWYEIARLDHPRPERGL--EQVITATSLRD-----DGGINV1 75
eNKeIcAdVkvGtvtqjEGeaseevLtadPaklk1KyavGvaSflqpGfdy
NK++cAD+-VGt+I++E dPak+k+Ky+GvaSflq+Gdd+
gi|1676758 76 -NKGYNPDF-EMWQKTEGKA---YFTGSFNRRAALKV---SFFGPFVGGY 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|pix|IA39486 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 195: score 600.2, E = 2.6e-175
-*+mkwMkLLLAlaLagvfgaRdrdsvgKcrvpssPRGfrfVknNFDv
mkw+++LLLaA + +aErd Crv+s frVknNFD+
gi|5803139 1 MEKWVALLLALLA-W--AAARD-----CRVSS---FRVKENFDK 45
KDPEGLFLGdnIvAFevDEKGMsaAtAGKrVRLnnNdvCADmGtFtD
KDPEGLFLGdnIvAFevDE+G+msaAtAGKrVRLnnNdvCADmGtFtD
gi|5803139 46 KDPEGLFLGdnIvAFevDE+G+msaAtAGKrVRLnnNdvCADmGtFtD 95
+EDPAKFKmKYNgvAsFlgkGndDHwiDtDygtAvgyScRllnLDGtC
+EDPAKFKmKYNgvAsFlgkGndDHwiDtDygtAvgyScRllnLDGtC
gi|5803139 96 TDDPAKFKmKYNgvAsFlgkGndDHwiVDTDYtYAVQYSScRllnLDGtC 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:

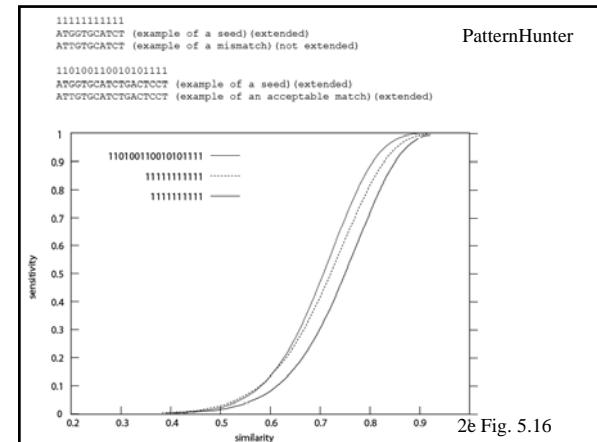
- There are exons (protein-coding sequence) and introns (intervening sequences).
- There may be sequencing errors or polymorphisms
- The comparison may between be related species (e.g. human and mouse)

BLAST-related tools for genomic DNA

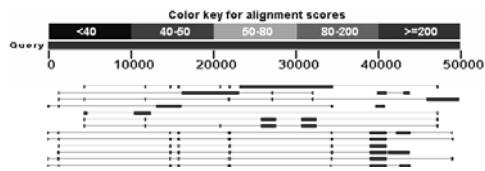
Recently developed tools include:

- 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



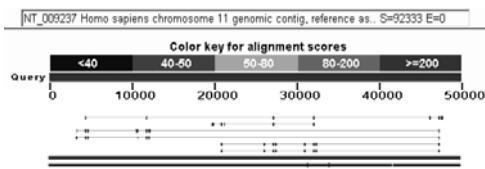
MegaBLAST at NCBI



MegaBLAST

Sequences producing significant alignments:	
(Click headers to sort column)	Description
M18208.1	Pongo pygmaeus gamma-1 and gamma-2 globin genes, or
M18209.1	Orangutan (P.pygmaeus) beta- and eta-globin pseudogene;
M18210.1	Orangutan delta-globin gene with Alu repeats in flanking
M18211.1	Orangutan beta- and delta-globin intergenic region w/
M18212.1	Orangutan delta-globin gene, complete cds
M18213.1	Orangutan gamma-2-fetal globin gene, complete cds
M18214.1	Orangutan gamma-1-fetal globin gene, complete cds
M18215.1	Orangutan gamma-1-globin gene, complete cds

MegaBLAST: 50 kilobases of the globin locus



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

UCSC Genome Bioinformatics

Genomes Gene Sorter BLAT PCR Tables FAQ Help

About the UCSC Genome Bioinformatics Site

This site contains the reference genome and working draft assemblies for a large collection of genomes. It also shows the CTFX (cytogenetic) browser, 13 species atlases, and a page for the ENCODE project.

Genome Browser Gene Sorter BLAT In Situ PCR Table Browser Utilities Downloads Release Log Custom Data ENCODE

New 10 September 2004 - Tetraodon Genome Assembly in Genome Browser

The Genome v7 Tetraodon nigroviridis genome assembly is now available in the UCSC Genome Browser and BLAT server. This assembly, UCSC version r6111, dated Feb. 2004, is the result of a collaboration between [Oncorhynchus](#) and the [Broad Institute](#) of MIT and Harvard.

"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

Human BLAT Search

BLAT Search Genome

Genome: Human Assembly: July 2003 Query type: BLAT's query Sort output: query.score Output type: hyperlink Submit

Paste DNA or protein sequence here in the FASTA format

```
g1|1840727|1|e1|NM_004744.2 Homo sapiens retinol binding protein 4, plasm
```

Only DNA sequences of 25,000 or fewer bases and proteins or translated sequence of 5000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total size for multiple sequence submissions is 50,000 bases or 12,500 letters.

About BLAT

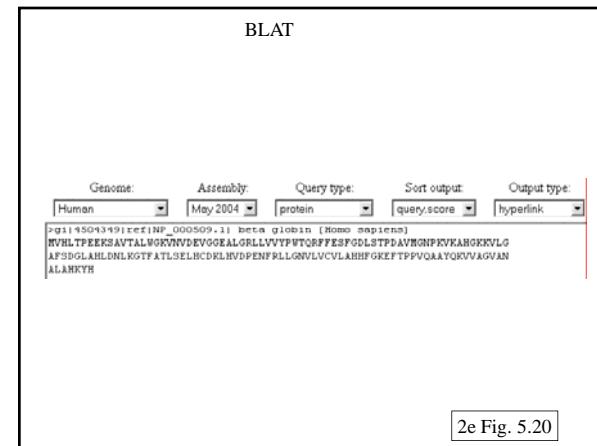
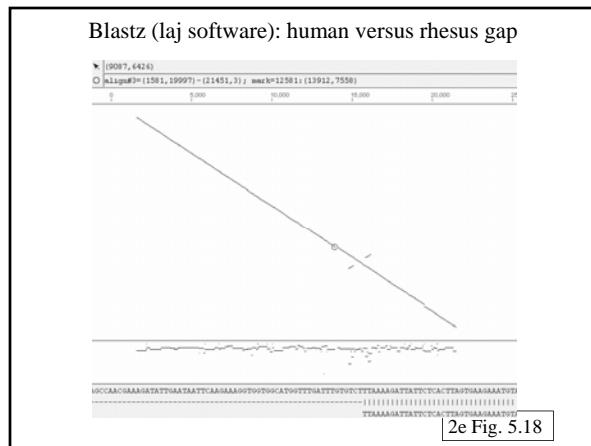
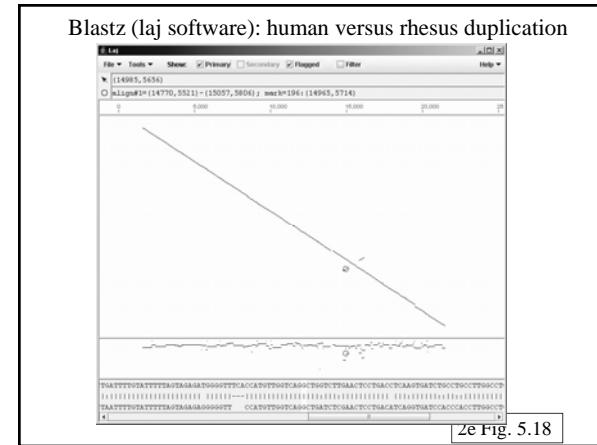
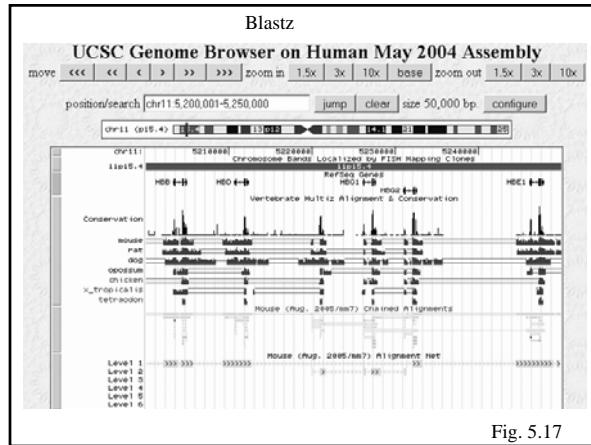
BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a lot less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a relatively modest Linux box. The index is used at the level of peptide matching, which is done outside memory for a detailed alignment.

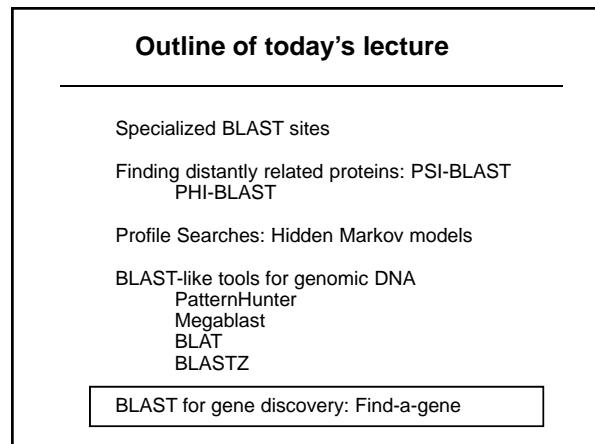
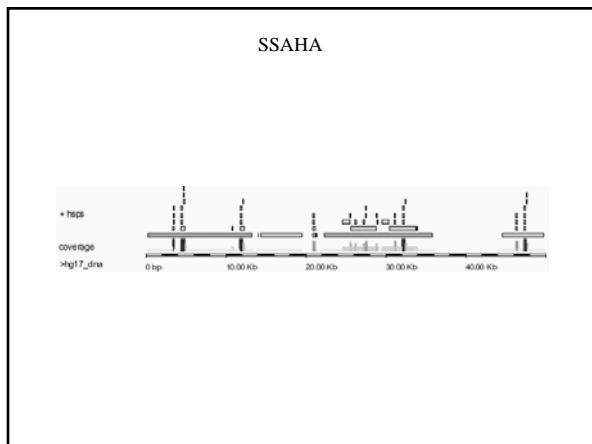
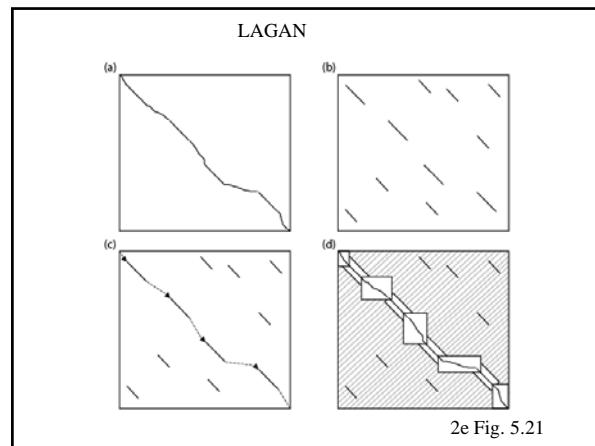
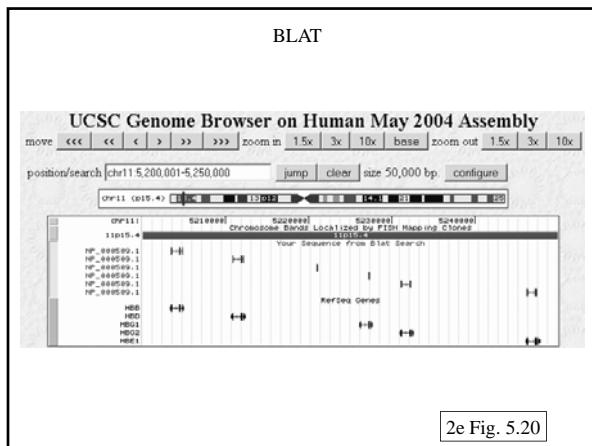
BLAT output includes browser and other formats

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QNAME	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	NM_004744.2	902	1	919	919	99.5%	10	-	95014180	95015504	9997
browser details	NM_004744.2	21	887	909	919	86.4%	9	-	77690017	77690038	22

UCSC Genome Browser on Human July 2003 Assembly





BLAST for gene discovery

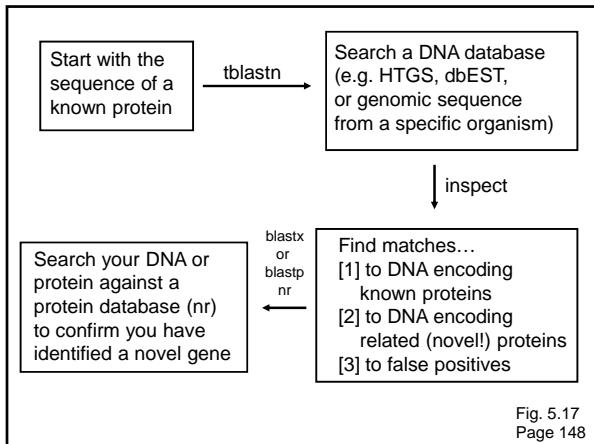
You can use BLAST to find a “novel” gene

Page 147

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.

Page 147

Find-a-gene project: HBB *Aquilegia*

```

>gb|1DT731130.1| EST1164900 Aquilegia cDNA library Aquilegia formosa x Aquilegia pubescens CDM clone C017402, mRNA sequence.
Length=679
Score = 38.1 bits (87), Expect = 0.20
Identities = 33/146 (22%), Positives = 63/146 (43%), Gaps = 11/146 (7%)
Frame = +2
Query 5 TPEKESVATALRG---KIVNDVVEKGEGALGRVLIVVTTUTDPRFFSGDLSTPOAVMGNPKV 62
  T +++++ V M K N E + + L + P + F D T + NPK
Sbjct 95 TEQEAQALVQE2MEIMQKQIPEL3QIUFFTTTLEIAFAAAKGSLF3YLD+TDEVYFQNNFLK 266
Query 63 ANHQQYLQAFEDSLAHL-----ENLKQTPATTSKELMCDNLWD/PENPFLLONVLCVLA 116
  AH KV + + L D + T L + H K +D P + F + + L + +
Sbjct 269 ANAHVVFPMTCCEAAVQLRENGA70FESTKLYGAHVWVGVIDP+HFEVYVKEALLTRIK 445
Query 117 HRHGRKFTPPQAGAYQVAVGVAL 142
  R + + + + + + + + + + + + + + + + + + + + + + +
Sbjct 446 DGVCENWSEELCIGAHEAHTPATAI 523
  
```

2e Fig. 5.23a

Find-a-gene project: *Aquilegia_Arabi*

```

>ref|NP_187663.1| UG ARR2 (NON-SYMBIOTIC HAEMOGLOBIN 2) [Arabidopsis thaliana]
sp|024521|HMB2_ARATH C Non-symbiotic hemoglobin 2 (Mb2) [ARATH_GLR2]
gb|IAA51381.1|AC011560.1| C class 2 non-symbiotic hemoglobin:69592-70841 [Arabidopsis thaliana]
gb|IAA862770.1| C class 2 non-symbiotic hemoglobin [Arabidopsis thaliana]
gb|IAA776353.1| C class 2 non-symbiotic hemoglobin [Arabidopsis thaliana]
gb|IAA65189.1| UG Non-symbiotic hemoglobin Mb2 [Arabidopsis thaliana]
gb|IAA749456.1| UG Atg10520 [Arabidopsis thaliana]
Length=158
Score = 243 bits (619), Expect = 1e-62
Identities = 116/154 (75%), Positives = 141/154 (91%), Gaps = 0/154 (0%)
Frame = +2
Query 02 ERVTEGGALV2E3WIMQUNPEAELSIQFTTILIAIAAAGAQLFSPFLMDTQDVYGNPKF 262
  E- FTE-GEALV2E3WEL-KQ+I+- SL FF+ LIIIAAAKGULFSPFLD+D+EVYVGNPKF
Sbjct 3 EIGFTEKQALVKEWSWILQKJIPKVYSLRFFSQIIFIAAAKGULFSFLRDSDEVYVGNPKF 62
Query 263 LKHAMAVCVPTCTCAAAVQLEBEGAVDILFESTLVYGAIVRYGQDIDPHYETVPAEALISTI 442
  LKHAMAVCVPTCTCAAAVQLEBEGAVDILFESTLVYGAIVRYGQDIDPHYETVPAEALISTI
Sbjct 63 LKHAMAVCVPTCTCAAAVQLEBEGAVDILFESTLVYGAIVRYGQDIDPHYETVPAEALISTI 122
Query 443 KEGVGEKNSXELCAGANSEKADYQLTAAIKTEKDKK 544
  KEGVGEKNSXELCAGANSEKADYQLTAAIKTEKDKK
Sbjct 123 KEGGLERTYNEEVEGANSQATDHLALAIKTEMKQ 156
  
```

2e Fig. 5.23b

Find-a-gene project: HBB *Vitis*

```

>emb|CAAP01000869.1| B Vitis vinifera, whole genome shotgun sequence of line PN40024,
contig 3451
Length=5962
Score = 31.2 bits (69), Expect = 9.2
Identities = 17/42 (40%), Positives = 24/42 (57%), Gaps = 1/42 (2%)
Frame = +3
Query 81 NLKGTFTAILSELHCDKLVHD/PENPFLLONVLCVLAQSFKE 122
  NLK T+ T S+ CD B + + + + + L H GKE
Sbjct 12864 NLK-TMTGDDSSD15M20LQKIQ1QDNPFLLILPFLKZ 12986
  
```

Find-a-gene project: myoglobin tblastn

```

>gb|1AAA02036297.1| B Oryza sativa (indica cultivar-group) Ctg036297, whole genome shotgun sequence
Length=20073
Score = 35.4 bits (80), Expect = 0.59
Identities = 50/57 (85%), Positives = 32/57 (56%), Gaps = 1/57 (1%)
Frame = +1
Query 22 IFCGNGQEVILIRLFGHGPETLKEFPLKFKWKLKSEDEMGAED-LQHGAGATVLTALGGIL 77
  +P H + +L+R G P TLE F P H + + + SE L+G+G+V A G +L
Sbjct 9040 VFHIFHFLQILLRHIGGRPLTLETFSLPCHGREGSMRLIRSEK*LEDGKSVRVAHGEVL 9210
  
```

2e Fig. 5.24a

Find-a-gene project: rice myoglobin blastp

```

>sp|P02193|MYG_DIDMA Myoglobin
Length=154
Score = 36.2 bits (82), Expect = 0.69, Method: Composition-based stats.
Identities = 20/57 (35%), Positives = 32/57 (56%), Gaps = 1/57 (1%)
Query 1 VVHHFQFLLPHIGGPPLVPLTLETFPLKFKWKLKSEDEMGAED-LQHGAGATVLTALGGIL 57
  +P H + +L+R G P TLE F P H + + + SE L+G+G+V A G +L
Sbjct 22 IFGHQGKVLIRLFGHGPETLKEFPLKFKWKLKSEDEMGAED-LQHGAGATVLTALGGIL 77
>sp|P02161|MYV_ZALCA Myoglobin
Length=154
Score = 35.8 bits (81), Expect = 0.78, Method: Composition-based stats.
Identities = 21/54 (38%), Positives = 32/54 (59%), Gaps = 1/54 (1%)
Query 4 HPLQILIPHIGGPPLVPLTLETFPLKFKWKLKSEDEMGAED-LQHGAGATVLTALGGIL 57
  H + +L+R G P TLE F P H + + + SE L+G+G+V A G +L
Sbjct 25 HQQEVLIDLPKFKWKLKSEDEMGAED-LQHGAGATVLTALGGIL 77
  
```

2e Fig. 5.24b

Microbial Genomes Blast Databases

BLAST with Microbial Genomes EBI Genomes Taxonomy NCBI NIH

Contributing Genome Centers: complete Microbial Genomes are available in [Eztrez Genomes](#). Infraspecific microbial genomes are not yet in GenBank and are not presented here. Click here or click on the CDS organism name below for more information about organisms whose sequences are contributing genome center. Click here for a table with information about all the databases represented here. [\[About this database\]](#)

CNGB: Please select one of the following databases to search

Genomic Center: Both All [114] Archaea [11] Eubacteria [96]
 Eukaryote [7] Other Input
[\[Click here to choose individual genomes to search\]](#)

Select the database to use and enter your sequence in the text area below.

Sequence: The query sequence is Program blast BLAST for low complexity regions by default.
 GTC
 HED
 MTE
 Submit Query Clear Input

Advanced options for the BLAST server:
 Expect Filter Descriptions Alignments
 Longer Queries [10] [default] [100] [1000]

BLAST with microbial genomes (13 bacterial 14 eukaryotic genomes tree)

Current results for BLAST searches are as requests from selected completed and unfinished genomes and partially sequenced. Partial genome sequences have been graciously provided by the sequencing centers or extracted from GenBank. See [\[About the Database\]](#) NCBI microbial sequencing centers to submit partially sequenced genomes to be included in the BLAST page. You can submit your data to us via ftp for which you will first need to contact us at [\[Support\]](#).

? indicates the ability to search against protein sequences. - completed genomic sequence, - unfinished genomic sequence, HED - submissions from GenBank, See [\[Help\]](#) for details.

Enter your query sequence as GI or FASTA:

Select type of query and database or BLAST PROGRAM:
 Query: DNA Database Protein BLAST program BLASTP
 Tool may change BLAST option:
 Local [14] Global [14] Tandemwise [14] Alignment [14]

Selected: **1-4.5e-5** genes (BLAST Adm-BLAST)

Show: alphabetical menu

Check if you want BL to select only completed genomes

BE Archaea
 BE Eucaryota
 BE Bacteria
 BE Bacteriophages
 BE Viruses
 BE Fungi
 BE Nematodes
 BE Protozoa
 BE Invertebrates
 BE Vertebrates
 BE Plants
 BE Fungi
 BE Archaeabacteria
 BE Halobacteria sp. NSC-1
 BE Methanococcus
 BE Methanomicrobacteria fermentoriphaga
 BE Methanococcales

Sequences producing significant alignments:

	Score	E	(bits)	Value
gnlISanger_6011S.typhi	Salmonella typhi CT18	Salmonella typhi A ...	37	0.13
gnlISanger_6011S.typhi	Salmonella typhi CT18	Salmonella paratyphi A ...	36	0.17
gnlISanger_6011S.typhi	Salmonella typhi	Salmonella typhimurium L ...	35	0.29
gnlISanger_6011S.typhi	Salmonella typhi	Salmonella typhimurium ...	35	0.29
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli O157:H7, complete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli O157:H7, incomplete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli K-12 MG1655 complete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli K-12 MG1655	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli K-12 MG1655	35	0.50
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	32	3.2
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	31	5.5
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	31	7.2
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	31	7.2
gnlISDSTC_54761C.albiciana	Contig=1956	size=11146 (bases 2 ...)	30	9.3
embLAL009126185UB	Bacillus subtilis	Bacillus subtilis complete genome	30	9.3
gnlISDSTC_54761C.albiciana	Contig=2286	size=21641 [ER01] [N...]	30	9.3

Alignments

```
>gnlISanger_6011S.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 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 66
      V NFD R= GTVY +4+ D +4+ A +4+ S+ G ++ EG V U
Sbjct: 4558772 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT- 4558596

Query: 87 ADMVGTFTDTEPDKFEEKFG 108
      + T+ + A R+ +G
Sbjct: 4558595 TEKGAYFTGDPSPRAALKYVSFFG 4558530
```

(Page 150)

Sequences producing significant alignments:

	Score	E	(bits)	Value
gnlISanger_6011S.typhi	Salmonella typhi CT18	Salmonella typhi A ...	37	0.13
gnlISanger_6011S.typhi	Salmonella typhi CT18	Salmonella paratyphi A ...	36	0.17
gnlISanger_6011S.typhi	Salmonella typhi	Salmonella typhimurium L ...	35	0.29
gnlISanger_6011S.typhi	Salmonella typhi	Salmonella typhimurium ...	35	0.29
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli O157:H7, complete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli O157:H7, incomplete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli K-12 MG1655 complete genome	35	0.50
gnlISanger_6011S.typhi	Escherichia coli	Escherichia coli K-12 MG1655	35	0.50
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	32	3.2
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	31	5.5
gnlITGR01581012	Treponema denticola	Treponema denticola unfinished frag...	31	7.2
gnlISDSTC_54761C.albiciana	Contig=1956	size=11146 (bases 2 ...)	30	9.3
embLAL009126185UB	Bacillus subtilis	Bacillus subtilis complete genome	30	9.3
gnlISDSTC_54761C.albiciana	Contig=2286	size=21641 [ER01] [N...]	30	9.3

Alignments

```
>gnlISanger_6011S.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 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 66
      V NFD R= GTVY +4+ D +4+ A +4+ S+ G ++ EG V U
Sbjct: 4558772 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT- 4558596

Query: 87 ADMVGTFTDTEPDKFEEKFG 108
      + T+ + A R+ +G
Sbjct: 4558595 TEKGAYFTGDPSPRAALKYVSFFG 4558530
```

this is a good candidate for a novel gene/protein

Alignments

```
>gi|4775481|emb|CAH42626_11 (Y17716) hypothetical protein [Klebsiella oxytoca]
Length = 177
Score = 161 bits (400), Expect = 9e-40
Identities = 75/81 (92%), Positives = 70/81 (95%)

Query: 1 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 60
      V NFD R= GTVY +4+ D +4+ A +4+ S+ G ++ EG V U
Sbjct: 30 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 60

Query: 61 EGGAYFTGDPSPRAALKYVSFFG 61
A blastp nr search confirms that the Salmonella query is closely related to other lipocalins
Sbjct: 90 EGGAYFTGDPSPRAALKYVSFFG 61

>gi|2497702|sp|Q4G031|BLC_CITPER OUTER MEMBRANE LIPOPROTEIN BLC PRECURSOR
21|21210191|gi|457010|outer membrane lipoprotein - Citrobacter freundii
21|717133|gb|ACI4556_11 (U21727) lipocalin precursor [Citrobacter freundii]
Length = 177

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

Query: 1 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 60
      V NFD R= GTVY +4+ D +4+ A +4+ S+ G ++ EG V U
Sbjct: 30 VVNNNDPDKPFGTVVAAKEDPPLDFGLPDKFEGLEQVTATTSLRDLDDGGINVKNGTNPAREWQKT 60

Query: 61 EGGAYFTGDPSPRAALKYVSFFG 61
      EGGAYFTG P+ AALEKVSSFFG
Sbjct: 90 EGGAYFTGDPSPRAALKYVSFFG 61
```

(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.