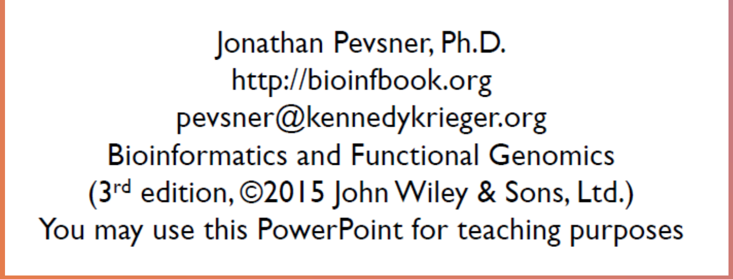
Chapter IV: Advanced Database Searching

Presentations use info from:



Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition, ©2015 John Wiley & Sons, Ltd.)
You may use this PowerPoint for teaching purposes

dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si

Local vs. Global Alignment

- **Local**

  -Suitable for aligning more divergent sequences or distantly related sequences.
  -Binds local regions with the highest level of similarity between the two sequences.
  -Used to finding out conserved patters in DNA sequences or conserved domains or motifs in two proteins

- **Global**

  -Suitable for aligning two closely related sequences.
  -An attempt is made to align the entire sequence (end to end alignment).
  -Usually done for comparing homologous genes like comparing two genes with the same function – or comparing two proteins with similar function.

What will you learn?

define a position-specific scoring matrix (PSSM);

explain how position-specific iterated BLAST (PSI-BLAST) and DELTA-BLAST greatly improve the sensitivity of BLAST protein searches;

describe profile hidden Markov models (HMMs) and explain their advantages over BLAST for database searching;

explain how spaced seed strategies improve the sensitivity of DNA searches; and

describe how millions of next-generation sequencing reads are aligned to a reference genome.

Outline

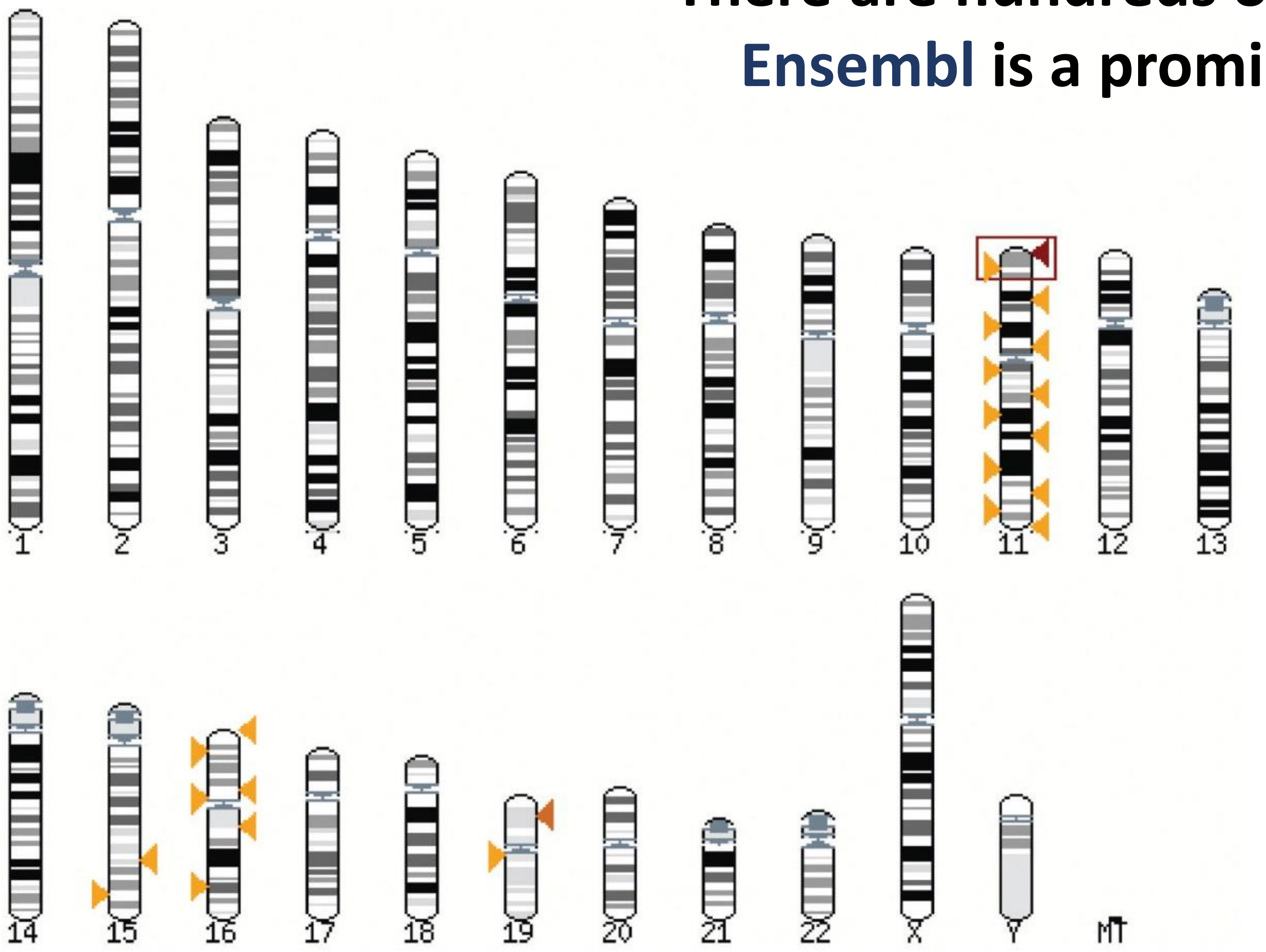
- Introduction
- Specialized BLAST sites
    - Organism-specific BLAST sites; specialized algorithms
- Finding distantly related proteins: PSI-BLAST) and DELTA-BLAST
    - Reverse Position-Specific BLAST
    - Domain enhanced lookup time Accelerated BLAST (DELTA-BLAST)
    - Assessing performance of PSI-BLAST and DELTA-BLAST
    - Pattern-hit initiated BLAST (PHI-BLAST)
- Profile searches: Hidden Markov Models and HMMER
- BLAST-like alignment tools to search genomic DNA
    - Benchmarking to assess genomic alignment performance
    - PatternHunter, BLASTZ, Enredo/Pecan, MegaBLAST, BLAT, LAGAN, SSAHA2
- Aligning NGS reads to a reference genome
    - Alignment based on hash tables; Burrows–Wheeler transform
- Perspective

Three problems standard BLAST cannot solve

1. Use human beta globin as a query against human RefSeq proteins, and BLASTP does not “find” human myoglobin. This is because the two proteins are too distantly related. PSI-BLAST at NCBI as well as hidden Markov models easily solve this problem.

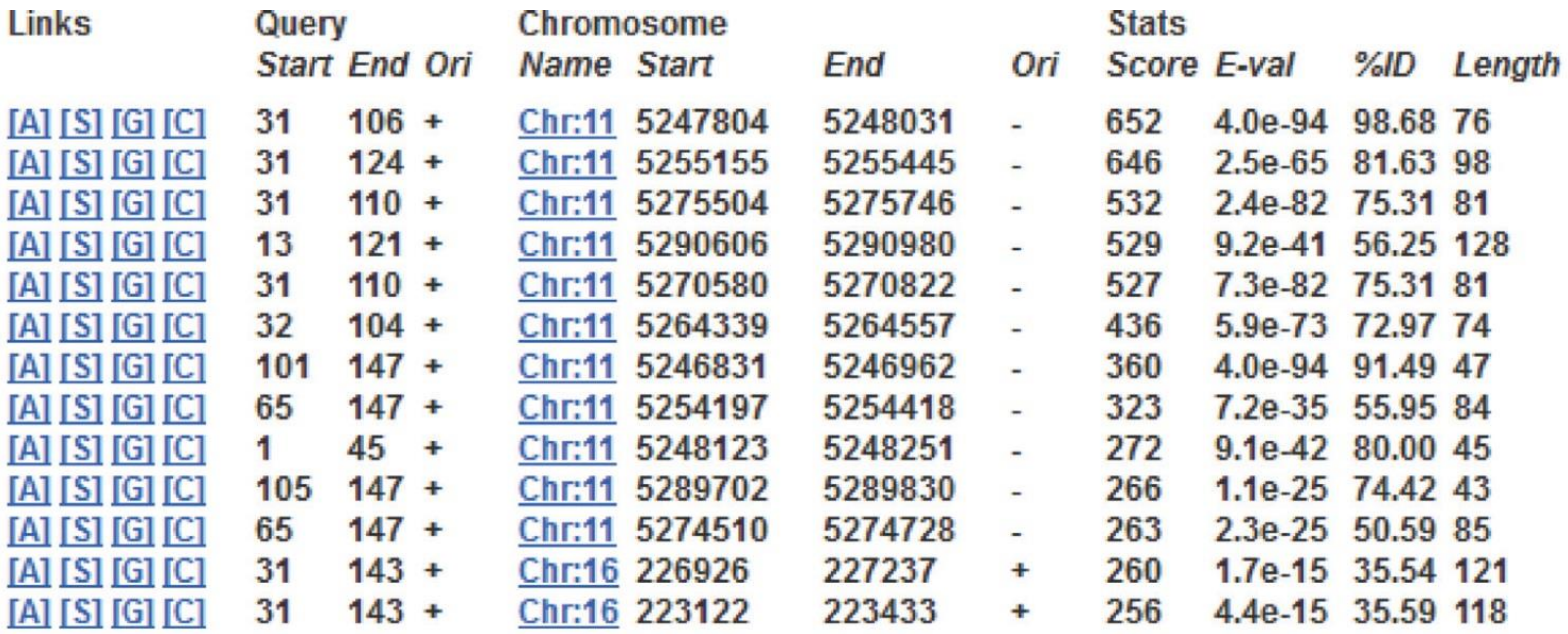
2. How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as MegaBLAST, BLAT, and LASTZ.

There are hundreds of BLAST servers; Ensembl is a prominent example



Ensembl can show database matches conveniently superimposed on an ideogram of chromosomes

There are hundreds of BLAST servers; Ensembl is a prominent example



Links	Query			Chromosome				Stats			
	Start	*End*	*Ori*	*Name*	*Start*	*End*	*Ori*	*Score*	*E-val*	*%ID*	*Length*
[A] [S] [G] [C]	31	106	+	Chr:11	5247804	5248031	-	652	4.0e-94	98.68	76
[A] [S] [G] [C]	31	124	+	Chr:11	5255155	5255445	-	646	2.5e-65	81.63	98
[A] [S] [G] [C]	31	110	+	Chr:11	5275504	5275746	-	532	2.4e-82	75.31	81
[A] [S] [G] [C]	13	121	+	Chr:11	5290606	5290980	-	529	9.2e-41	56.25	128
[A] [S] [G] [C]	31	110	+	Chr:11	5270580	5270822	-	527	7.3e-82	75.31	81
[A] [S] [G] [C]	32	104	+	Chr:11	5264339	5264557	-	436	5.9e-73	72.97	74
[A] [S] [G] [C]	101	147	+	Chr:11	5246831	5246962	-	360	4.0e-94	91.49	47
[A] [S] [G] [C]	65	147	+	Chr:11	5254197	5254418	-	323	7.2e-35	55.95	84
[A] [S] [G] [C]	1	45	+	Chr:11	5248123	5248251	-	272	9.1e-42	80.00	45
[A] [S] [G] [C]	105	147	+	Chr:11	5289702	5289830	-	266	1.1e-25	74.42	43
[A] [S] [G] [C]	65	147	+	Chr:11	5274510	5274728	-	263	2.3e-25	50.59	85
[A] [S] [G] [C]	31	143	+	Chr:16	226926	227237	+	260	1.7e-15	35.54	121
[A] [S] [G] [C]	31	143	+	Chr:16	223122	223433	+	256	4.4e-15	35.59	118

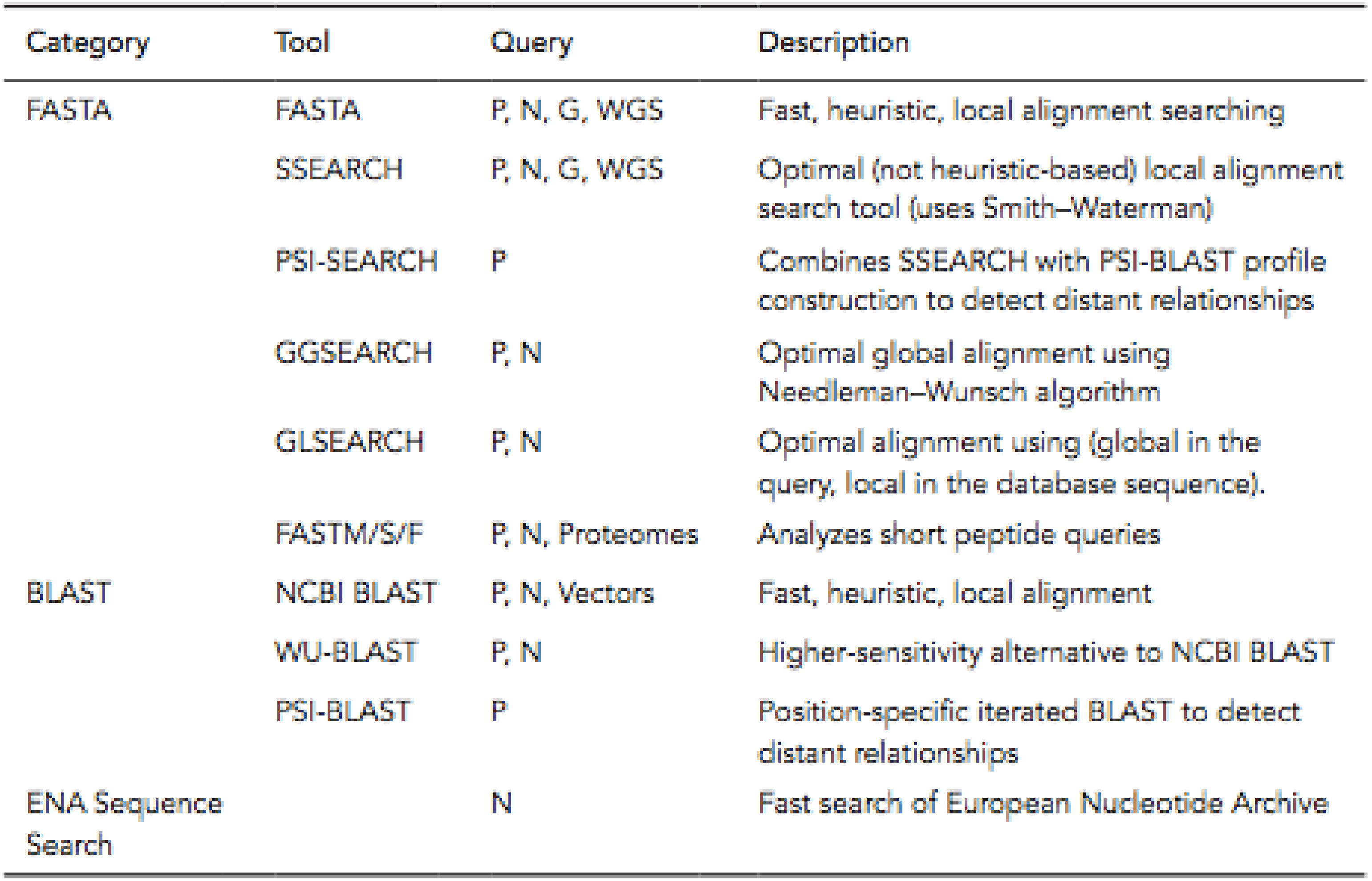
Ensembl BLAST output summarizes scores, expect values and other features.

Specialized BLAST-related algorithms

There are numerous specialized BLAST-related algorithms

BLAST of next-generation sequence (NGS) data

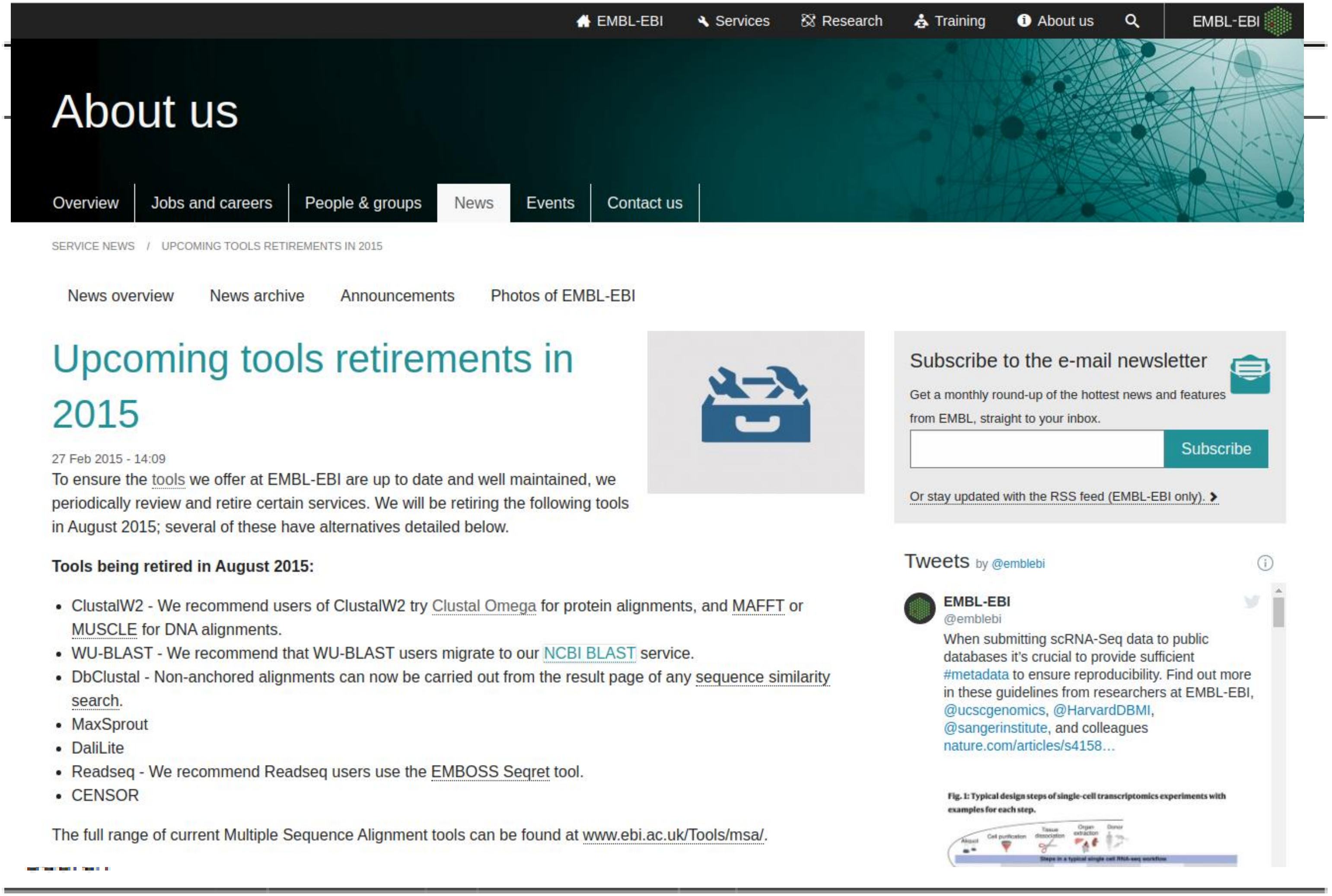
Sequence similarity searching tools at EBI



Category	Tool	Query	Description
FASTA	FASTA	P, N, G, WGS	Fast, heuristic, local alignment searching
	SSEARCH	P, N, G, WGS	Optimal (not heuristic-based) local alignment search tool (uses Smith–Waterman)
	PSI-SEARCH	P	Combines SSEARCH with PSI-BLAST profile construction to detect distant relationships
	GGSEARCH	P, N	Optimal global alignment using Needleman–Wunsch algorithm
	GLSEARCH	P, N	Optimal alignment using (global in the query, local in the database sequence).
	FASTM/S/F	P, N, Proteomes	Analyzes short peptide queries
BLAST	NCBI BLAST	P, N, Vectors	Fast, heuristic, local alignment
	WU-BLAST	P, N	Higher-sensitivity alternative to NCBI BLAST
	PSI-BLAST	P	Position-specific iterated BLAST to detect distant relationships
ENA Sequence Search		N	Fast search of European Nucleotide Archive

P, protein; N, nucleotide; G, genomes; WGS, whole-genome shotgun

Sequence similarity searching tools at EBI



Upcoming tools retirements in 2015

27 Feb 2015 - 14:09

To ensure the tools we offer at EMBL-EBI are up to date and well maintained, we periodically review and retire certain services. We will be retiring the following tools in August 2015; several of these have alternatives detailed below.

Tools being retired in August 2015:

- ClustalW2 - We recommend users of ClustalW2 try Clustal Omega for protein alignments, and MAFFT or MUSCLE for DNA alignments.
- WU-BLAST - We recommend that WU-BLAST users migrate to our NCBI BLAST service.
- DbClustal - Non-anchored alignments can now be carried out from the result page of any sequence similarity search.
- MaxSprout
- DaliLite
- Readseq - We recommend Readseq users use the EMBOSS Seqret tool.
- CENSOR

The full range of current Multiple Sequence Alignment tools can be found at www.ebi.ac.uk/Tools/msa/.

Subscribe to the e-mail newsletter

Get a monthly round-up of the hottest news and features from EMBL, straight to your inbox.

Subscribe

Or stay updated with the RSS feed (EMBL-EBI only).

Tweets by @emblebi

EMBL-EBI
@emblebi
When submitting scRNA-Seq data to public databases it's crucial to provide sufficient #metadata to ensure reproducibility. Find out more in these guidelines from researchers at EMBL-EBI, @ucscgenomics, @HarvardDBMI, @sangerinstitute, and colleagues nature.com/articles/s4158…

Fig. 1: Typical design steps of single-cell transcriptomics experiments with examples for each step.

P, protein; N, nucleotide; G, genomes; WGS, whole-genome shotgun

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.

PSI-BLAST is performed in five steps

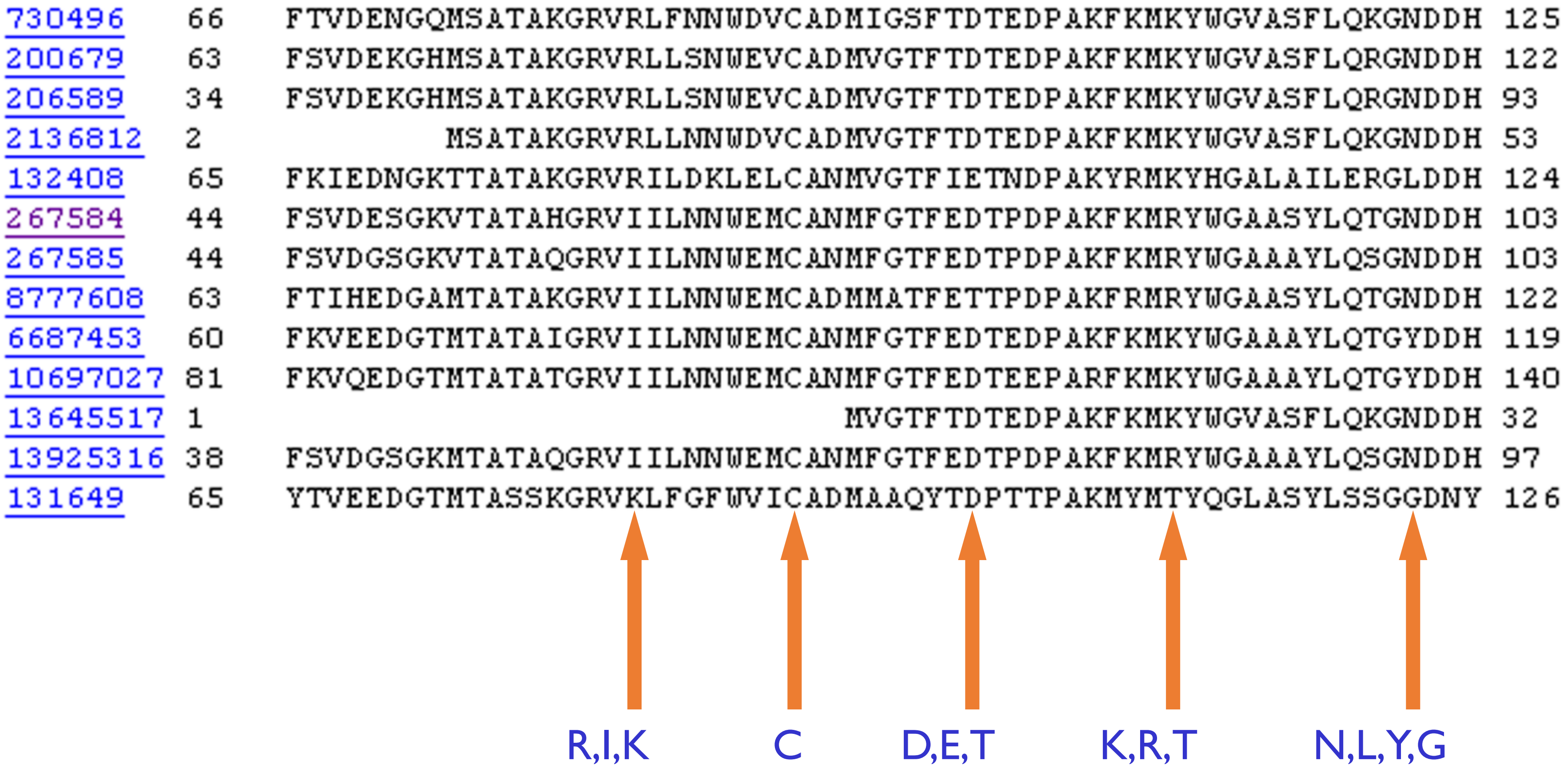
[1] Select a query and search it against a protein database

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)

Inspect the BLASTP output to identify empirical “rules” regarding amino acids tolerated at each position



```
730496   66  FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679   63  FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589   34  FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812  2           MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408   65  FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584   44  FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267585   44  FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103
8777608  63  FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453  60  FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027 81  FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517 1                               MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316 38  FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649   65  YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
```

R,I,K

C

D,E,T

K,R,T

N,L,Y,G

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-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

20 amino acids

all the amino acids from position 1 to the end of your PSI-BLAST query protein

```
         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
```

		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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-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

note that a given amino acid (such as alanine) in your query protein can receive different scores for matching alanine—depending on the position in the protein

		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										3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4										0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4										0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2										-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2										-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4												-3	-2	7	0	0
14	A	3	-2	-1	-2										-3	1	1	-1	-3	-3	-1
15	A	2	-1	0	-1										-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2										-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1										-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2										-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-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

note that a given amino acid (such as tryptophan) in your query protein can receive different scores for matching tryptophan—depending on the position in the protein

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

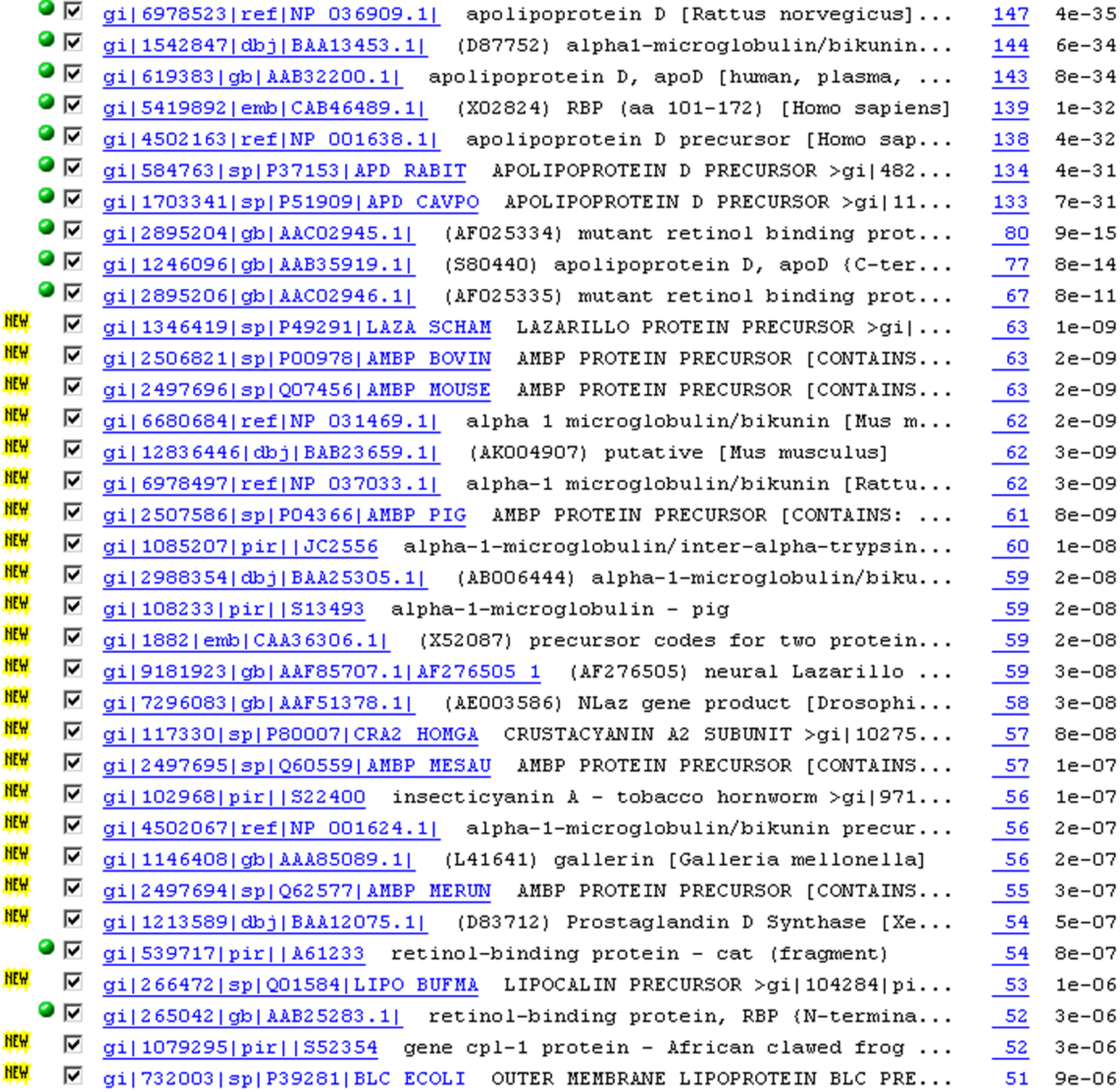
PSI-BLAST is performed in five steps

[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)



```
    ● ☑ gi|6978523|ref|NP_036909.1|  apolipoprotein D [Rattus norvegicus]...  147  4e-35
    ● ☑ gi|1542847|dbj|BAA13453.1|  (D87752) alpha1-microglobulin/bikunin...  144  6e-34
    ● ☑ gi|619383|gb|AAB32200.1|  apolipoprotein D, apoD [human, plasma, ...  143  8e-34
    ● ☑ gi|5419892|emb|CAB46489.1|  (X02824) 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|P37153|APD_RABIT  APOLIPOPROTEIN D PRECURSOR >gi|482...  134  4e-31
    ● ☑ gi|1703341|sp|P51909|APD_CAVPO  APOLIPOPROTEIN D PRECURSOR >gi|11...  133  7e-31
    ● ☑ gi|2895204|gb|AAC02945.1|  (AF025334) mutant retinol binding prot...   80  9e-15
    ● ☑ gi|1246096|gb|AAB35919.1|  (S80440) apolipoprotein D, apoD (C-ter...   77  8e-14
    ● ☑ gi|2895206|gb|AAC02946.1|  (AF025335) mutant retinol binding prot...   67  8e-11
NEW   ☑ gi|1346419|sp|P49291|LAZA_SCHAM  LAZARILLO PROTEIN PRECURSOR >gi|...   63  1e-09
NEW   ☑ gi|2506821|sp|P00978|AMBP_BOVIN  AMBP PROTEIN PRECURSOR [CONTAINS...   63  2e-09
NEW   ☑ gi|2497696|sp|Q07456|AMBP_MOUSE  AMBP PROTEIN PRECURSOR [CONTAINS...   63  2e-09
NEW   ☑ gi|6680684|ref|NP_031469.1|  alpha 1 microglobulin/bikunin [Mus m...   62  2e-09
NEW   ☑ gi|12836446|dbj|BAB23659.1|  (AK004907) putative [Mus musculus]        62  3e-09
NEW   ☑ gi|6978497|ref|NP_037033.1|  alpha-1 microglobulin/bikunin [Rattu...   62  3e-09
NEW   ☑ gi|2507586|sp|P04366|AMBP_PIG  AMBP PROTEIN PRECURSOR [CONTAINS: ...   61  8e-09
NEW   ☑ gi|1085207|pir||JC2556  alpha-1-microglobulin/inter-alpha-trypsin...   60  1e-08
NEW   ☑ gi|2988354|dbj|BAA25305.1|  (AB006444) alpha-1-microglobulin/biku...   59  2e-08
NEW   ☑ gi|108233|pir||S13493  alpha-1-microglobulin - pig                     59  2e-08
NEW   ☑ gi|1882|emb|CAA36306.1|  (X52087) precursor codes for two protein...   59  2e-08
NEW   ☑ gi|9181923|gb|AAF85707.1|AF276505_1  (AF276505) neural Lazarillo ...   59  3e-08
NEW   ☑ gi|7296083|gb|AAF51378.1|  (AE003586) NLaz gene product [Drosophi...   58  3e-08
NEW   ☑ gi|117330|sp|P80007|CRA2_HOMGA  CRUSTACYANIN A2 SUBUNIT >gi|10275...   57  8e-08
NEW   ☑ gi|2497695|sp|Q60559|AMBP_MESAU  AMBP PROTEIN PRECURSOR [CONTAINS...   57  1e-07
NEW   ☑ gi|102968|pir||S22400  insecticyanin A - tobacco hornworm >gi|971...   56  1e-07
NEW   ☑ gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin precur...   56  2e-07
NEW   ☑ gi|1146408|gb|AAA85089.1|  (L41641) gallerin [Galleria mellonella]     56  2e-07
NEW   ☑ gi|2497694|sp|Q62577|AMBP_MERUN  AMBP PROTEIN PRECURSOR [CONTAINS...   55  3e-07
NEW   ☑ gi|1213589|dbj|BAA12075.1|  (D83712) Prostaglandin D Synthase [Xe...   54  5e-07
    ● ☑ gi|539717|pir||A61233  retinol-binding protein - cat (fragment)        54  8e-07
NEW   ☑ gi|266472|sp|Q01584|LIPO_BUFMA  LIPOCALIN PRECURSOR >gi|104284|pi...   53  1e-06
    ● ☑ gi|265042|gb|AAB25283.1|  retinol-binding protein, RBP (N-termina...   52  3e-06
NEW   ☑ gi|1079295|pir||S52354  gene cpl-1 protein - African clawed frog ...   52  3e-06
NEW   ☑ gi|732003|sp|P39281|BLC_ECOLI  OUTER MEMBRANE LIPOPROTEIN BLC PRE...   51  9e-06
```

PSI-BLAST is performed in five steps

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

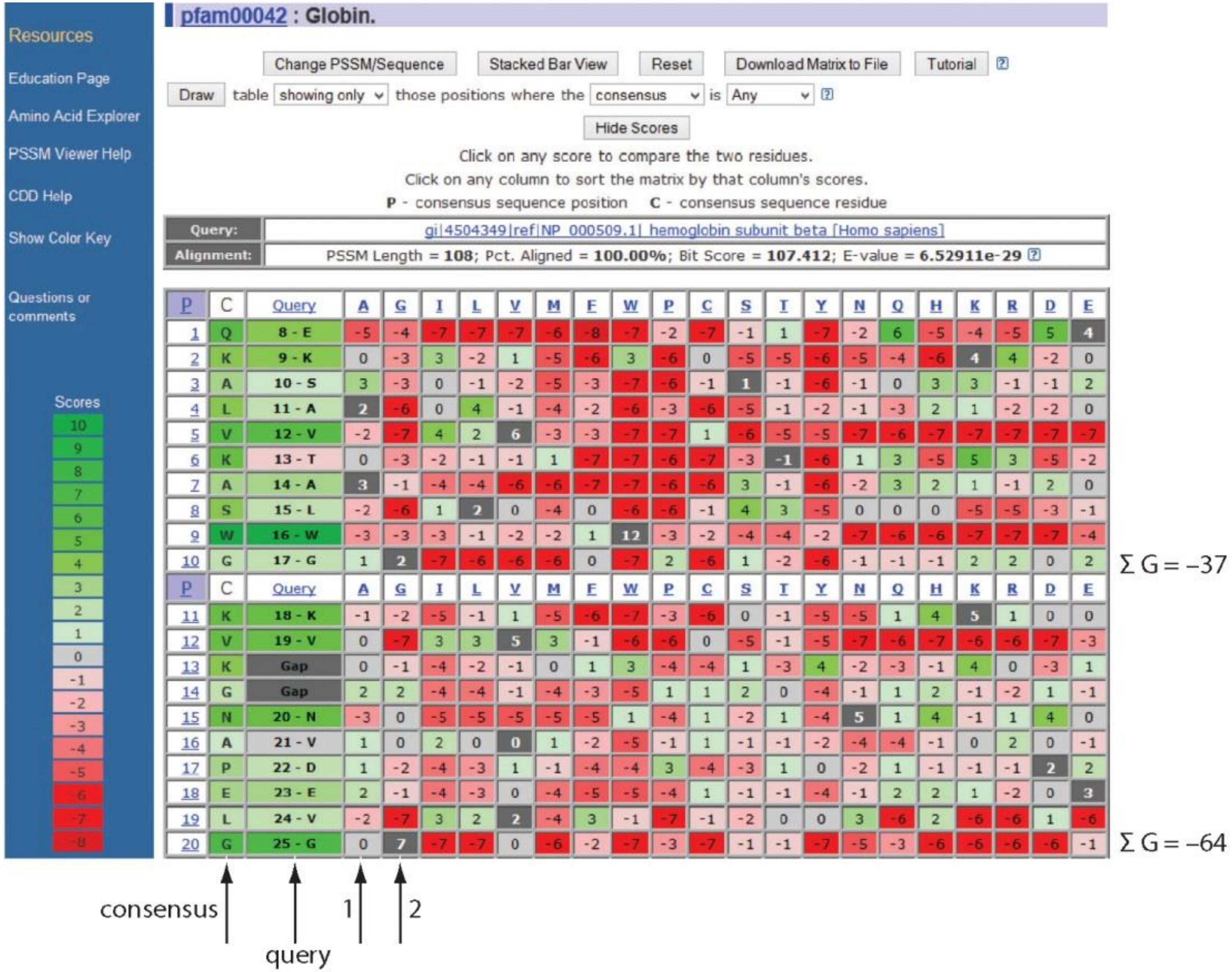
[2] PSI-BLAST constructs a multiple sequence alignment
then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[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.

Position-specific scoring matrix (PSSM)



pfam00042 : Globin.

Draw table showing only those positions where the consensus is Any

Click on any score to compare the two residues.
Click on any column to sort the matrix by that column's scores.
P - consensus sequence position **C** - consensus sequence residue

Query:	gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
Alignment:	PSSM Length = **108**; Pct. Aligned = **100.00%**; Bit Score = **107.412**; E-value = **6.52911e-29**

P	C	Query	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	Q	8 - E	-5	-4	-7	-7	-7	-6	-8	-7	-2	-7	-1	1	-7	-2	6	-5	-4	-5	5	4
2	K	9 - K	0	-3	3	-2	1	-5	-6	3	-6	0	-5	-5	-6	-5	-4	-6	4	4	-2	0
3	A	10 - S	3	-3	0	-1	-2	-5	-3	-7	-6	-1	1	-1	-6	-1	0	3	3	-1	-1	2
4	L	11 - A	2	-6	0	4	-1	-4	-2	-6	-3	-6	-5	-1	-2	-1	-3	2	1	-2	-2	0
5	V	12 - V	-2	-7	4	2	6	-3	-3	-7	-7	1	-6	-5	-5	-7	-6	-7	-7	-7	-7	-7
6	K	13 - T	0	-3	-2	-1	-1	1	-7	-7	-6	-7	-3	-1	-6	1	3	-5	5	3	-5	-2
7	A	14 - A	3	-1	-4	-4	-6	-6	-7	-7	-6	-6	3	-1	-6	-2	3	2	1	-1	2	0
8	S	15 - L	-2	-6	1	2	0	-4	0	-6	-6	-1	4	3	-5	0	0	0	-5	-5	-3	-1
9	W	16 - W	-3	-3	-3	-1	-2	-2	1	12	-3	-2	-4	-4	-2	-7	-6	-6	-7	-7	-7	-4
10	G	17 - G	1	2	-7	-6	-6	-6	0	-7	2	-6	1	-2	-6	-1	-1	-1	2	2	0	2
P	C	Query	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	K	18 - K	-1	-2	-5	-1	1	-5	-6	-7	-3	-6	0	-1	-5	-5	1	4	5	1	0	0
12	V	19 - V	0	-7	3	3	5	3	-1	-6	-6	0	-5	-1	-5	-7	-6	-7	-6	-6	-7	-3
13	K	Gap	0	-1	-4	-2	-1	0	1	3	-4	-4	1	-3	4	-2	-3	-1	4	0	-3	1
14	G	Gap	2	2	-4	-4	-1	-4	-3	-5	1	1	2	0	-4	-1	1	2	-1	-2	1	-1
15	N	20 - N	-3	0	-5	-5	-5	-5	-5	1	-4	1	-2	1	-4	5	1	4	-1	1	4	0
16	A	21 - V	1	0	2	0	0	1	-2	-5	-1	1	-1	-1	-2	-4	-4	-1	0	2	0	-1
17	P	22 - D	1	-2	-4	-3	1	-1	-4	-4	3	-4	-3	1	0	-2	1	-1	-1	-1	2	2
18	E	23 - E	2	-1	-4	-3	0	-4	-5	-5	-4	1	-1	-1	-4	-1	2	2	1	-2	0	3
19	L	24 - V	-2	-7	3	2	2	-4	3	-1	-7	-1	-2	0	0	3	-6	2	-6	-6	1	-6
20	G	25 - G	0	7	-7	-7	0	-6	-2	-7	-3	-7	-1	-1	-7	-5	-3	-6	-6	-6	-6	-1

Try it yourself

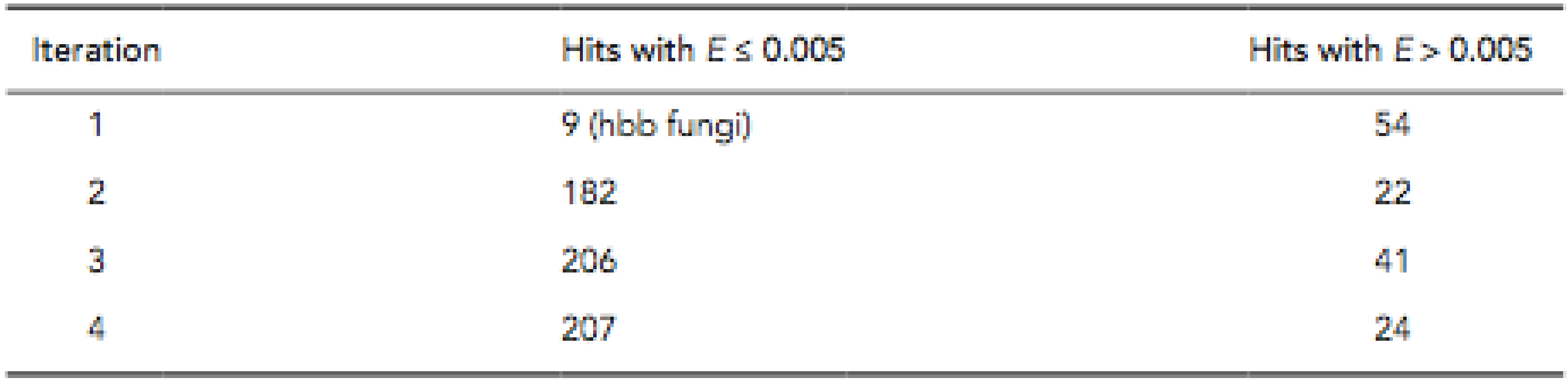
❑ Do a protein BLAST of accession NP_000509 against the Refseq_protein database and limit it to Fungi/Metazoa group.

❑ Repeat the analysis selecting the PSI-BLAST option.

❑ Perform the same analysis on the EMBLEBI website.

❑ Compare the results.

PSI-BLAST: dramatic increase in number of hits



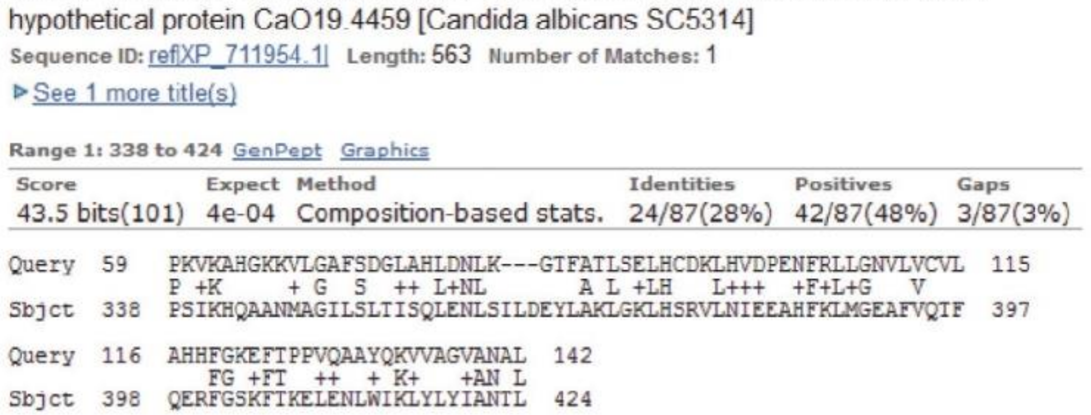
Iteration	Hits with $E \leq 0.005$	Hits with $E > 0.005$
1	9 (hbb fungi)	54
2	182	22
3	206	41
4	207	24

Given this query, a standard BLASTP search would produce about 9 hits with low expect values. This PSI-BLAST search produces >200 hits after 3 or 4 iterations.

Note that PSI-BLAST E values can improve dramatically!

After 1st iteration:
Expect = 4e-04
Alignment length = 87 amino acids

(a) PSI-BLAST iteration 1 match (human beta globin versus a *C. albicans* globin)



hypothetical protein CaO19.4459 [Candida albicans SC5314]

Sequence ID: ref|XP_711954.1| Length: 563 Number of Matches: 1

▶ See 1 more title(s)

Range 1: 338 to 424 GenPept Graphics

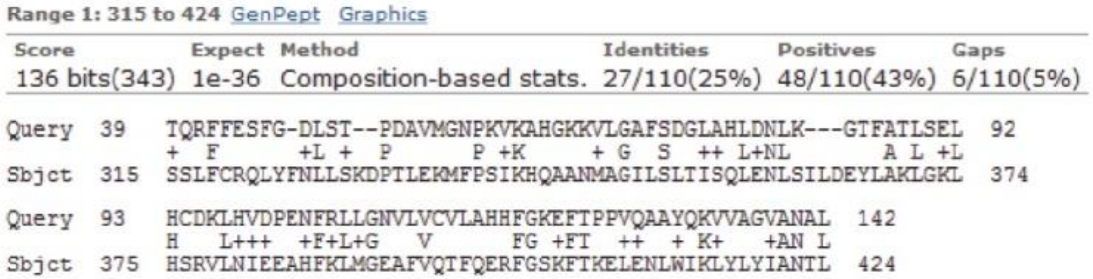
Score	Expect	Method	Identities	Positives	Gaps
43.5 bits(101)	4e-04	Composition-based stats.	24/87(28%)	42/87(48%)	3/87(3%)

```
Query  59   PKVKAHGKKVLGAFSDGLAHLDNLK---GTFATLSELHCDKLHVDPENFRLLGNVLVCVL  115
            P +K     + G  S  ++ L+NL       A L +LH   L+++  +F+L+G   V
Sbjct  338  PSIKHQAANMAGILSLTISQLENLSILDEYLAKLGKLHSRVLNIEEAHFKLMGEAFVQTF  397

Query  116  AHHFGKEFTPPVQAAYQKVVAGVANAL  142
               FG +FT  ++  + K+   +AN L
Sbjct  398  QERFGSKFTKELENLWIKLYLYIANTL  424
```

After 2nd iteration:
Expect = 1e-36
Alignment length = 110 amino acids

(b) PSI-BLAST iteration 2 (human beta globin versus a *C. albicans* globin)



Range 1: 315 to 424 GenPept Graphics

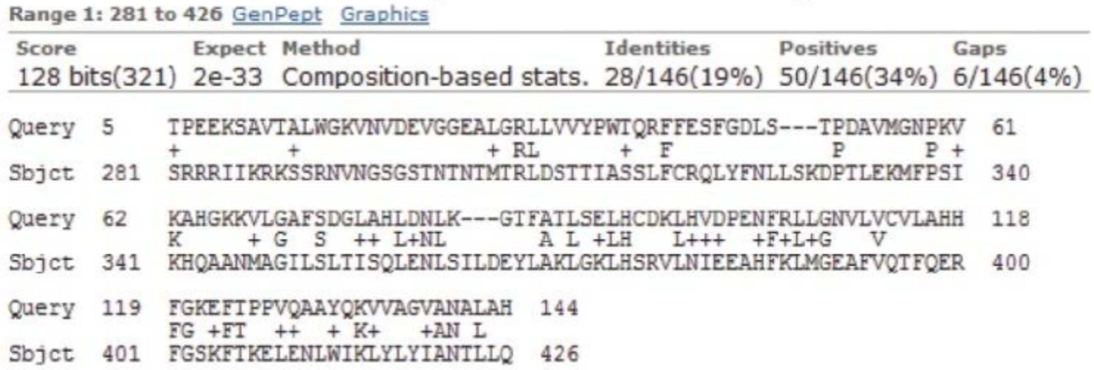
Score	Expect	Method	Identities	Positives	Gaps
136 bits(343)	1e-36	Composition-based stats.	27/110(25%)	48/110(43%)	6/110(5%)

```
Query  39   TQRFFESFG-DLST--PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLK---GTFATLSEL  92
            +  F      +L +  P      P +K     + G  S  ++ L+NL       A L +L
Sbjct  315  SSLFCRQLYFNLLSKDPTLEKMFPSIKHQAANMAGILSLTISQLENLSILDEYLAKLGKL  374

Query  93   HCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL  142
            H   L+++  +F+L+G   V      FG +FT  ++  + K+   +AN L
Sbjct  375  HSRVLNIEEAHFKLMGEAFVQTFQERFGSKFTKELENLWIKLYLYIANTL  424
```

After 3rd iteration:
Expect = 2e-33
Alignment length = 146 amino acids

(c) PSI-BLAST iteration 3 (human beta globin versus a *C. albicans* globin)



Range 1: 281 to 426 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
128 bits(321)	2e-33	Composition-based stats.	28/146(19%)	50/146(34%)	6/146(4%)

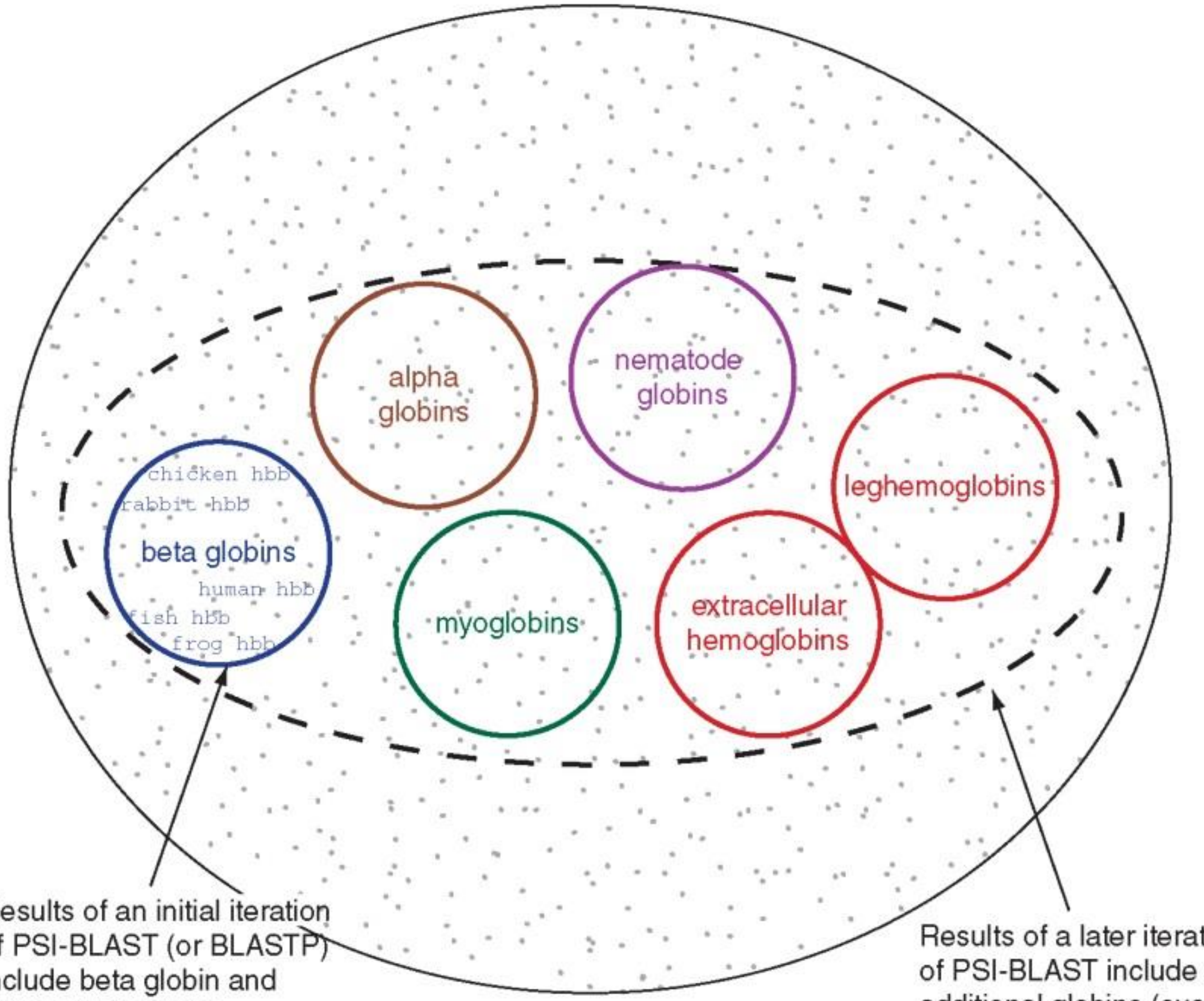
```
Query  5    TPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS---TPDAVMGNPKV  61
            +        +              + RL      +  F            P      P +
Sbjct  281  SRRRIIKRKSSRNVNGSGSTNTNTMTRLDSTTIASSLFCRQLYFNLLSKDPTLEKMFPSI  340

Query  62   KAHGKKVLGAFSDGLAHLDNLK---GTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH  118
            K     + G  S  ++ L+NL       A L +LH   L+++  +F+L+G   V
Sbjct  341  KHQAANMAGILSLTISQLENLSILDEYLAKLGKLHSRVLNIEEAHFKLMGEAFVQTFQER  400

Query  119  FGKEFTPPVQAAYQKVVAGVANALAH  144
            FG +FT  ++  + K+   +AN L
Sbjct  401  FGSKFTKELENLWIKLYLYIANTLLQ  426
```

PSI-BLAST algorithm increases the sensitivity of a database search by detecting homologous matches with relatively low sequence identity

All globins
(four main groups: globins, bacterial-like globins, protoglobins, phycobilisomes)



Results of an initial iteration of PSI-BLAST (or BLASTP) include beta globin and some other globins.

Results of a later iteration of PSI-BLAST include many additional globins (such as leghemoglobins) that were not detected initially. All bind heme and transport ligands such as oxygen.

PSI-BLAST: the problem of corruption

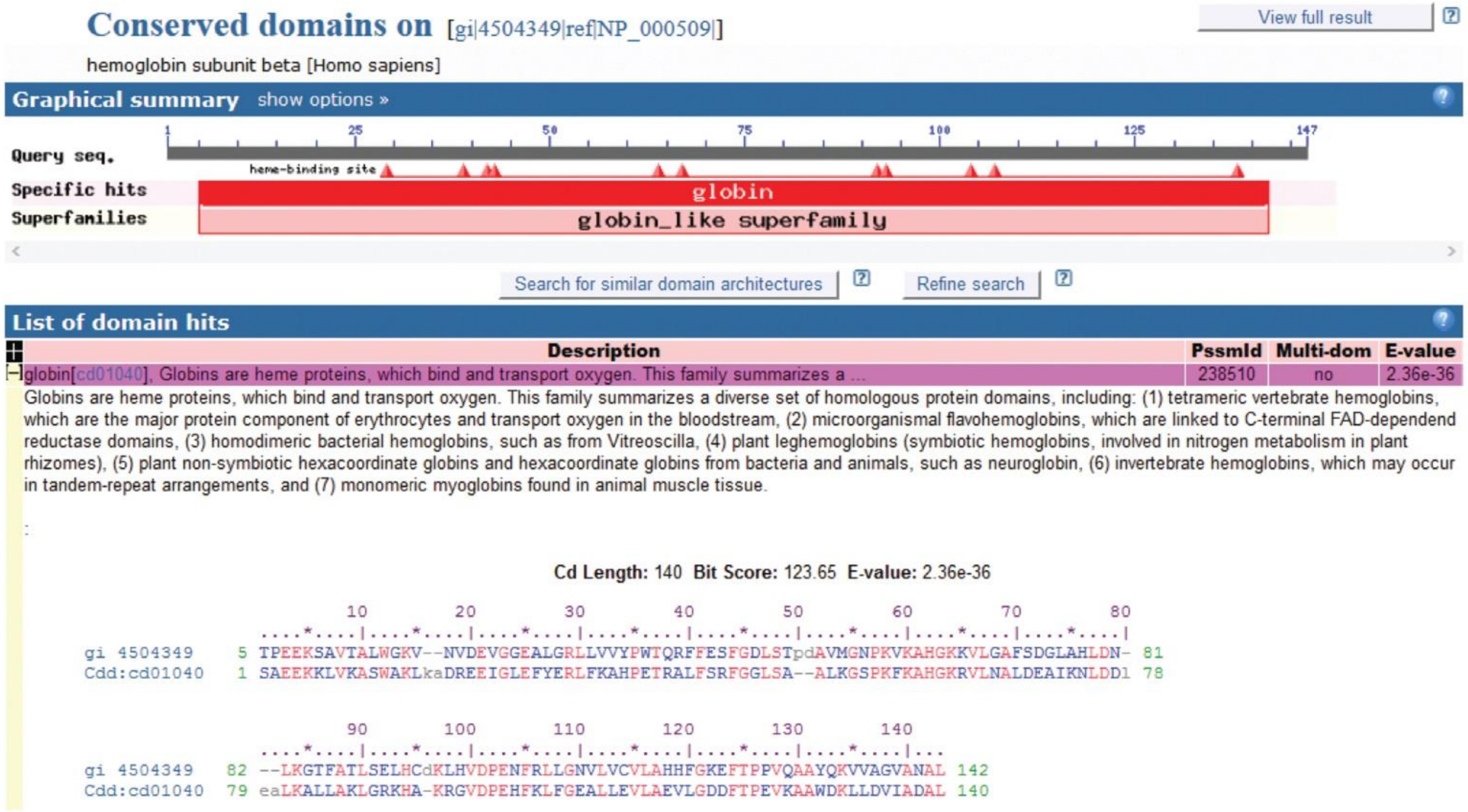
PSI-BLAST - once a match is incorporated into a PSSM it will never be removed, even if it is wrong (i.e. even if it is a false positive that is not truly homologous to the query).

Not only will it stay, but it may also lead to the inclusion of many other related false positive hits.

There are three main approaches to removing false positives:

(1) Filter biased amino acid regions. (This is an option in BLAST.)
(2) Lower the expect value threshold to make the search more stringent.
(3) Visually inspect the output from each PSI-BLAST iteration and remove suspicious matches (by unchecking the corresponding boxes).

Reverse position-specific BLAST (RPS-BLAST): search a query against a collection of predefined position-specific scoring matrices



Conserved domains on [gi|4504349|ref|NP_000509|]

hemoglobin subunit beta [Homo sapiens]

Description	PssmId	Multi-dom	E-value
globin[cd01040], Globins are heme proteins, which bind and transport oxygen. This family summarizes a ...	238510	no	2.36e-36

Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.

Cd Length: 140 **Bit Score:** 123.65 **E-value:** 2.36e-36

```
                           10        20        30        40        50        60        70        80
                   ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 4504349      5 TPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTpdAVMGNPKVKAHGKKVLGAFSDGLAHLDN- 81
Cdd:cd01040     1 SAEEKKLVKASWAKLkaDREEIGLEFYERLFKAHPETRALFSRFGGLSA--ALKGSPKFKAHGKRVLNALDEAIKNLDDl 78

                           90       100       110       120       130       140
                   ....*....|....*....|....*....|....*....|....*....|....*....|...
gi 4504349     82 --LKGTFATLSELHCdKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL 142
Cdd:cd01040    79 eaLKALLAKLGRKHA-KRGVDPEHFKLFGEALLEVLAEVLGDDFTPEVKAAWDKLLDVIADAL 140
```

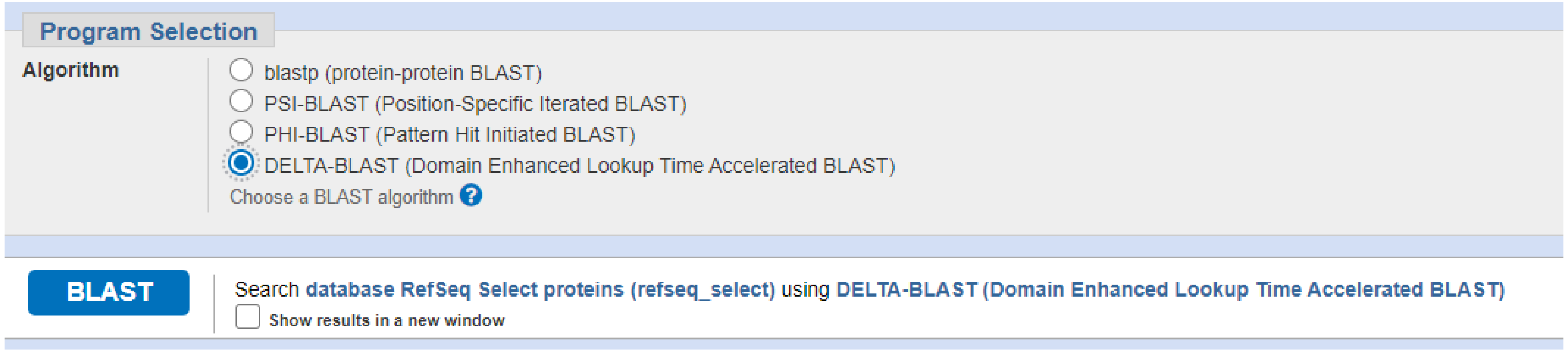
RPS-BLAST searches are incorporated into the Conserved Domain Database (CDD) at NCBI

DELTA-BLAST: better than PSI-BLAST!

In 2012 NCBI introduced DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST) to the family of BLASTP tools.

DELTA-BLAST constructs a PSSM using the results of a Conserved Domain Database (CDD) search and uses that to search a sequence database.

<u>The results are typically superior to those of PSI-BLAST</u>.



Program Selection

Algorithm

○ blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
● DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm ?

Search database RefSeq Select proteins (refseq_select) using DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

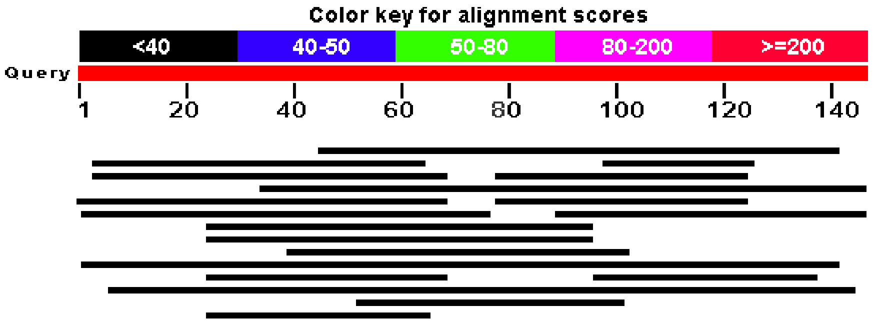
□ Show results in a new window

DELTA-BLAST: better than PSI-BLAST

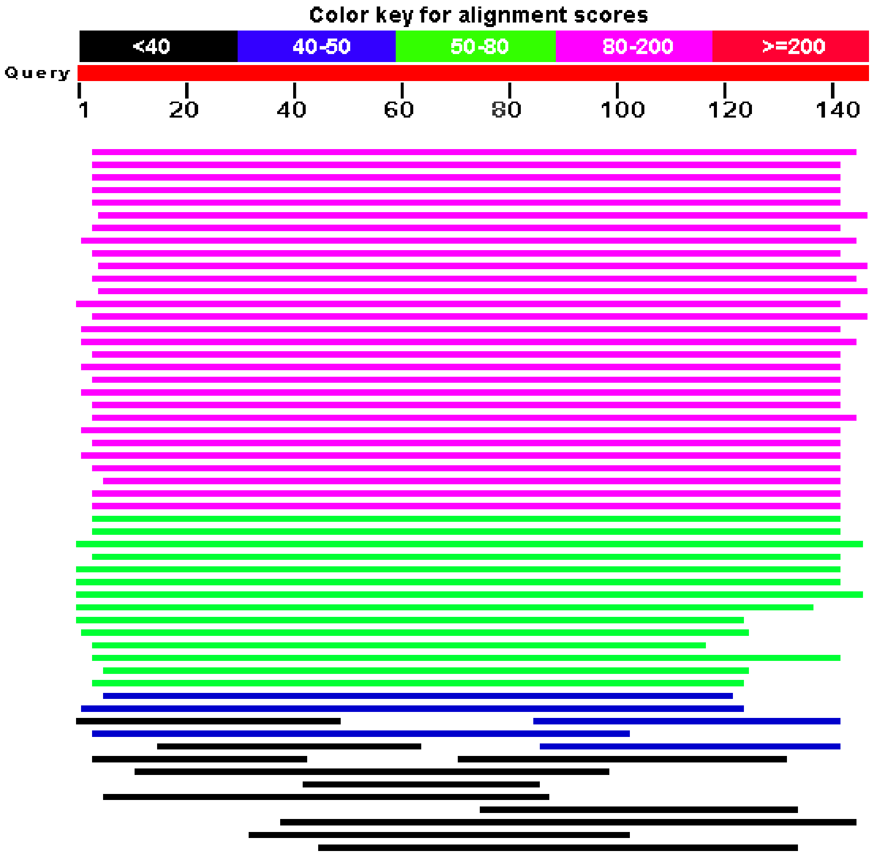
- Domain enhanced lookup time Accelerated BLAST (DELTA-BLAST) is faster, more sensitive and accurate than PSI-BLAST.
- PSI-BLAST creates multiple alignments and position-specific scoring matrices (PSSMs).
- DELTA-BLAST searches a query against a library of pre-computed PSSMs. One reason DELTA-BLAST outperforms PSI-BLAST is that it results in larger, more complete PSSMs than PSI-BLAST.
- Most queries do match a PSSM; if not the search proceeds in a PSI-BLAST-like manner.
- One iteration of DELTA-BLAST is recommended.

Search HBB (NP_000509) against RefSeq plants…

BLAST



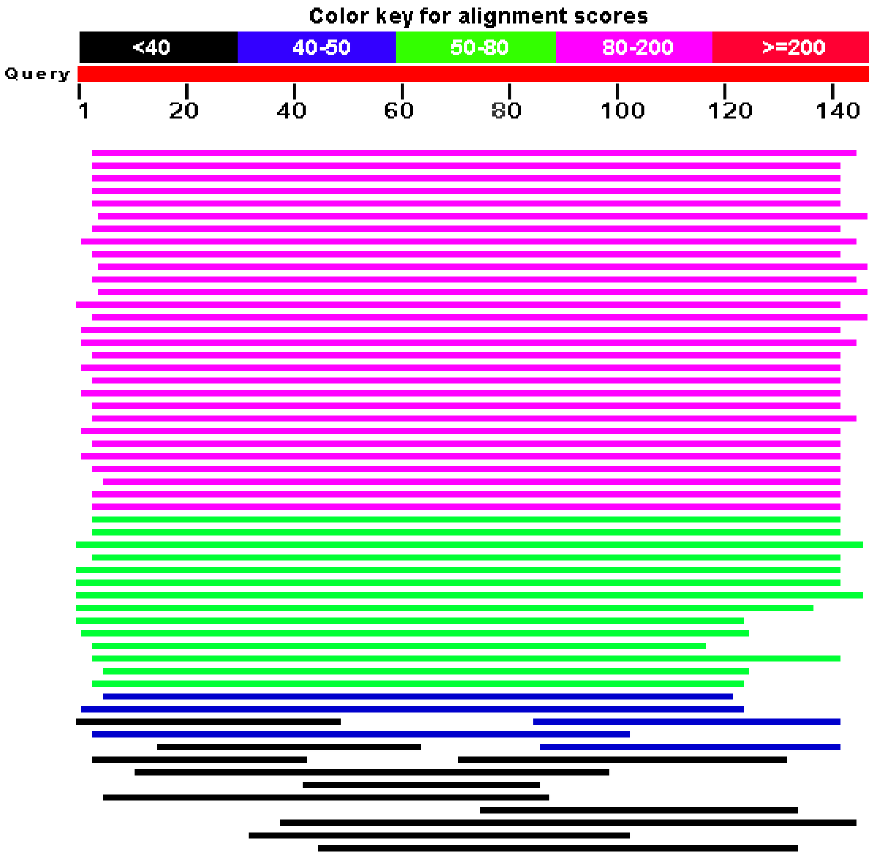
DELTA-BLAST



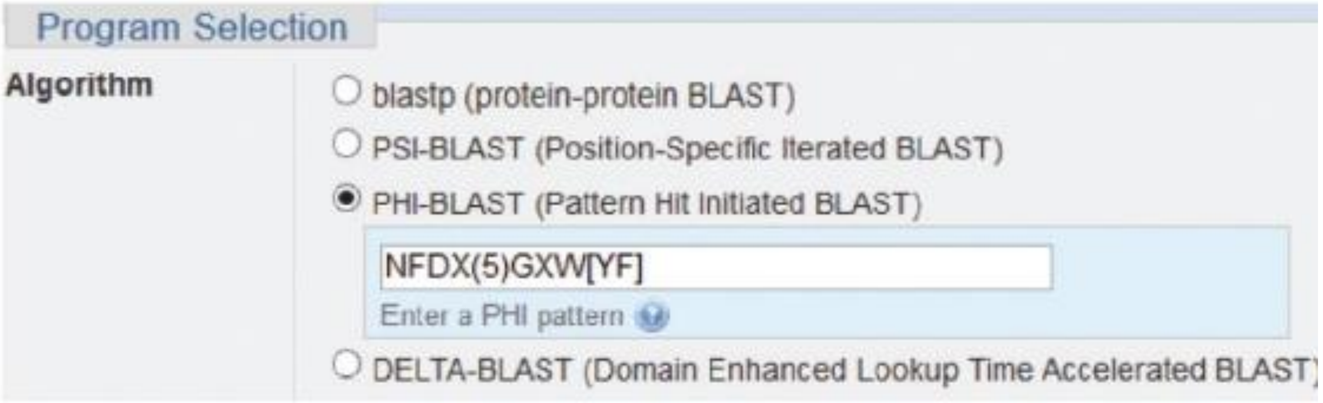
DELTA-BLAST

DELTA-BLAST is better than PSI-BLAST because it takes advantage of longer PSSMs.

If your query does not match any PSSM, DELTA-BLAST simply returns a BLASTP-like result.



PHI-BLAST: Pattern hit initiated BLAST



Program Selection

Algorithm

○ blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

● PHI-BLAST (Pattern Hit Initiated BLAST)

NFDX(5)GXW[YF]

Enter a PHI pattern

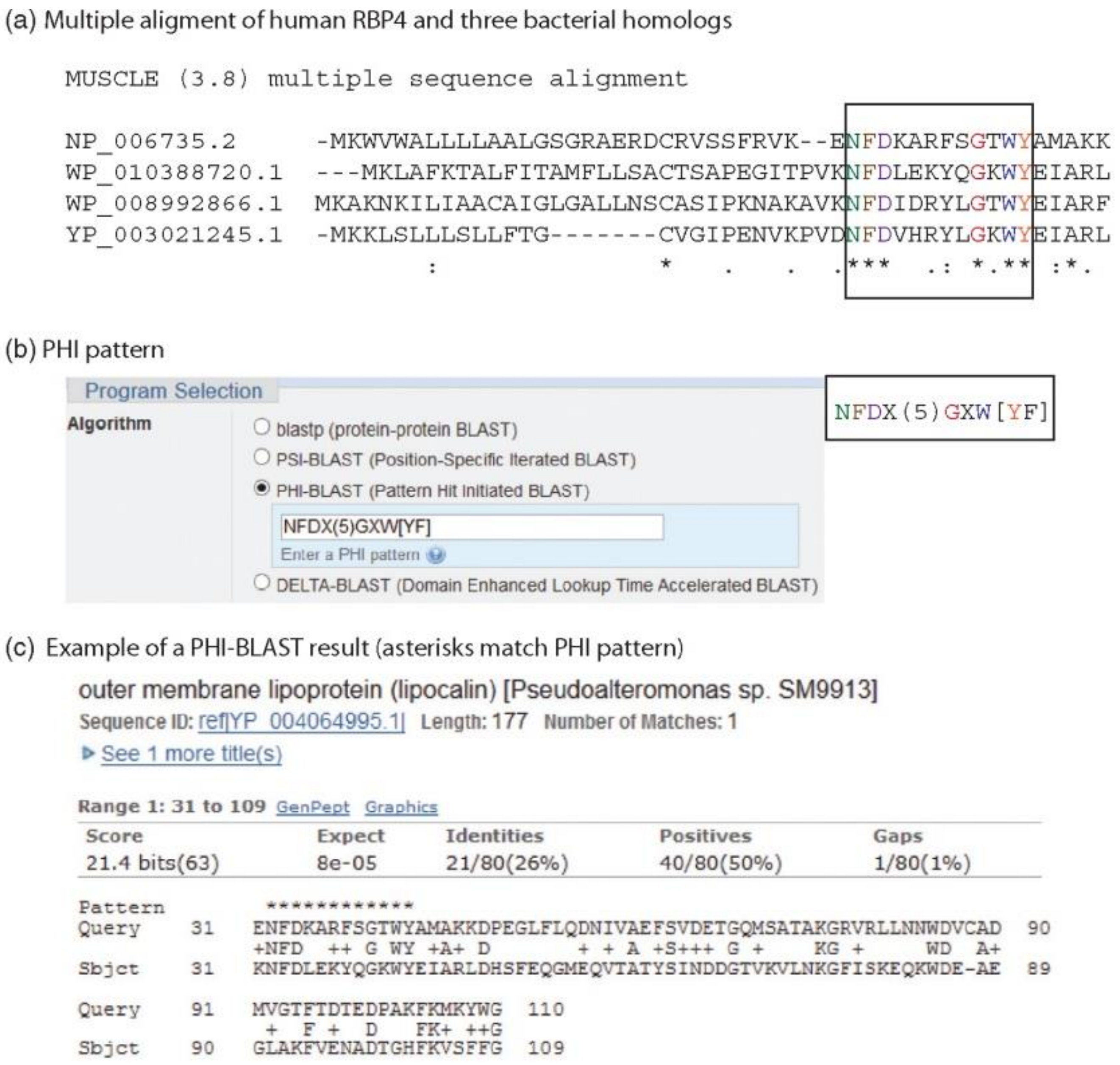
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Sometimes you have a protein query that has a known pattern. You can use PHI-BLAST to include that pattern, which can be user-selected or obtained from a database of such patterns such as PROSITE.

All resulting database matches must include that pattern (which is indicated with asterisks *** in the output).

PHI-BLAST is specialized and is not commonly used but can be very useful.

Choosing a pattern and performing a PHI-BLAST search



(a) Multiple aligment of human RBP4 and three bacterial homologs

```
MUSCLE (3.8) multiple sequence alignment

NP_006735.2     -MKWVWALLLLAALGSGRAERDCRVSSFRVK--ENFDKARFSGTWYAMAKK
WP_010388720.1  ---MKLAFKTALFITAMFLLSACTSAPEGITPVKNFDLEKYQGKWYEIARL
WP_008992866.1  MKAKNKILIAACAIGLGALLNSCASIPKNAKAVKNFDIDRYLGTWYEIARF
YP_003021245.1  -MKKLSLLLSLLFTG-------CVGIPENVKPVDNFDVHRYLGKWYEIARL
                       :              *   .   .  .***  .: *.** :*.
```

(b) PHI pattern

Program Selection

Algorithm

○ blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

● PHI-BLAST (Pattern Hit Initiated BLAST)

NFDX(5)GXW[YF]

Enter a PHI pattern

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

NFDX(5)GXW[YF]

(c) Example of a PHI-BLAST result (asterisks match PHI pattern)

outer membrane lipoprotein (lipocalin) [Pseudoalteromonas sp. SM9913]

Sequence ID: ref|YP_004064995.1| Length: 177 Number of Matches: 1

▶ See 1 more title(s)

Range 1: 31 to 109 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
21.4 bits(63)	8e-05	21/80(26%)	40/80(50%)	1/80(1%)

```
Pattern           ************
Query    31   ENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCAD  90
              +NFD  ++ G WY +A+ D       + + A +S+++ G +    KG +     WD  A+
Sbjct    31   KNFDLEKYQGKWYEIARLDHSFEQGMEQVTATYSINDDGTVKVLNKGFISKEQKWDE-AE  89

Query    91   MVGTFTDTEDPAKFKMKYWG  110
               +  F +   D   FK+ ++G
Sbjct    90   GLAKFVENADTGHFKVSFFG  109
```

Inspect an alignment, choose a pattern (manually).

Follow the rules for the syntax of your pattern.

The output includes asterisks indicating the position of your pattern.

Try it to boost sensitivity of your search.

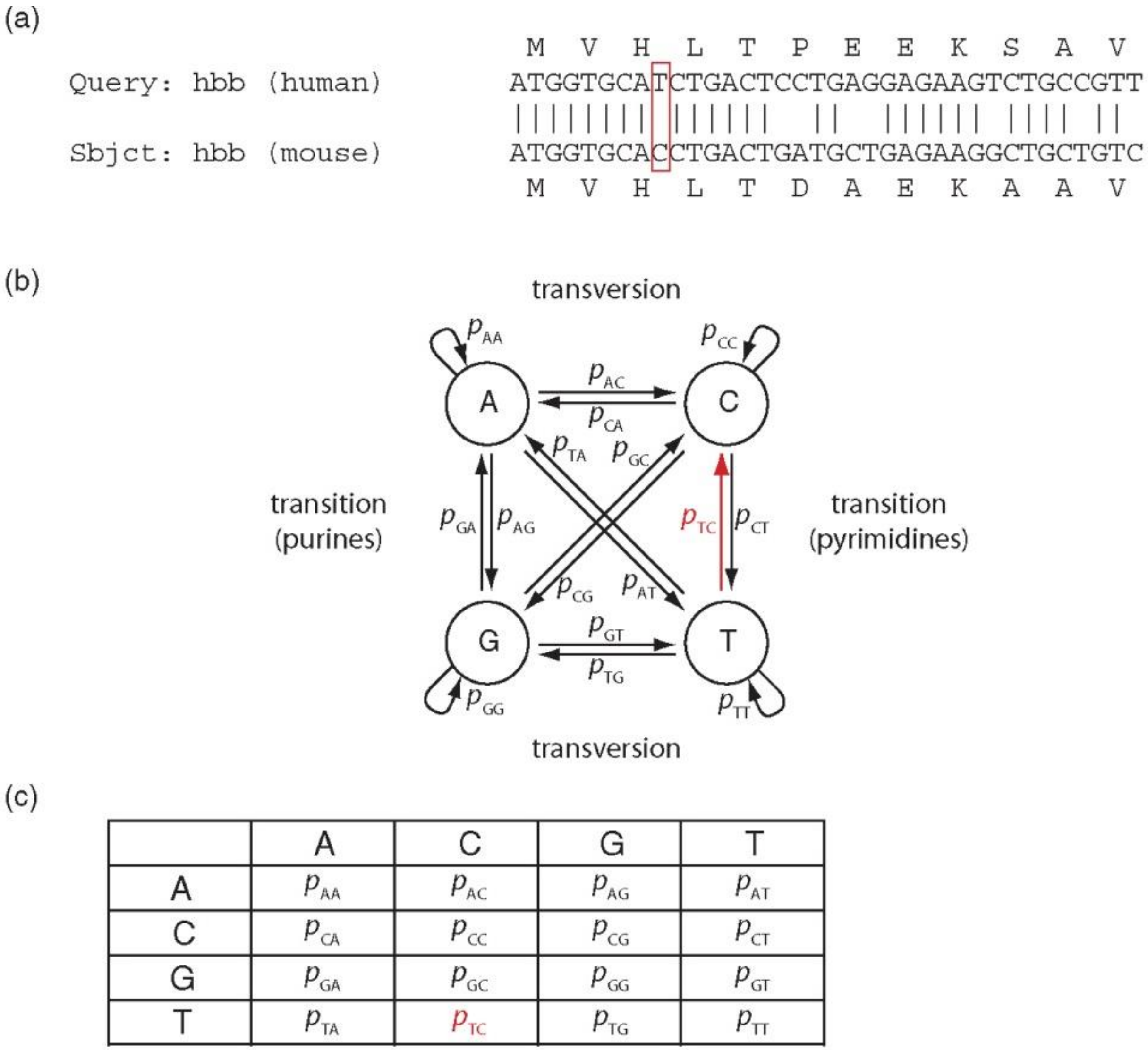
Multiple sequence alignment to profile HMMs

- In the 1990’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 (unlike DELTA-BLAST and PSI-BLAST)

A hidden Markov model describes the **transition probabilities for the alignment** of nucleotides (shown here) or amino acids



(a)

```
                     M  V  H  L  T  P  E  E  K  S  A  V
Query: hbb (human)  ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTT
                    |||||||| ||||||  ||  |||||| |||| ||
Sbjct: hbb (mouse)  ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTC
                     M  V  H  L  T  D  A  E  K  A  A  V
```

(b)

(c)

	A	C	G	T
A	p_{AA}	p_{AC}	p_{AG}	p_{AT}
C	p_{CA}	p_{CC}	p_{CG}	p_{CT}
G	p_{GA}	p_{GC}	p_{GG}	p_{GT}
T	p_{TA}	p_{TC}	p_{TG}	p_{TT}

Consider five globin protein segments (each consisting of five amino acids).

1D8U	HAMSV
1OJ6A	HIRKV
2hhbB	HGKKV
1FSL	HAEKL
2MM1	HGATV

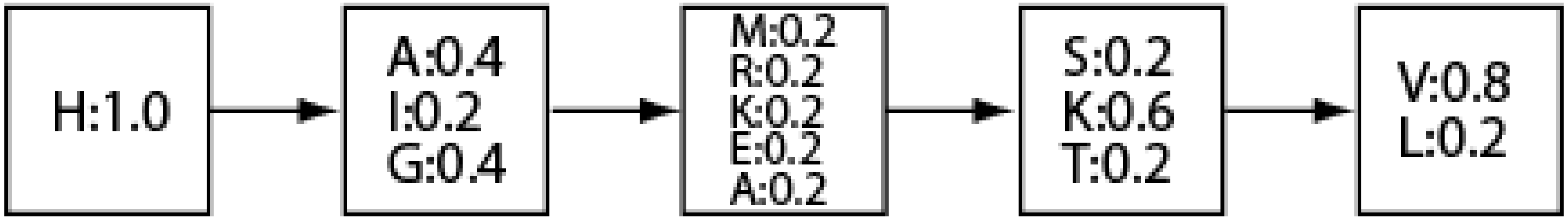
We can describe the probability of occurrence of an amino acid at each position.

	position				
Probability	1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
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

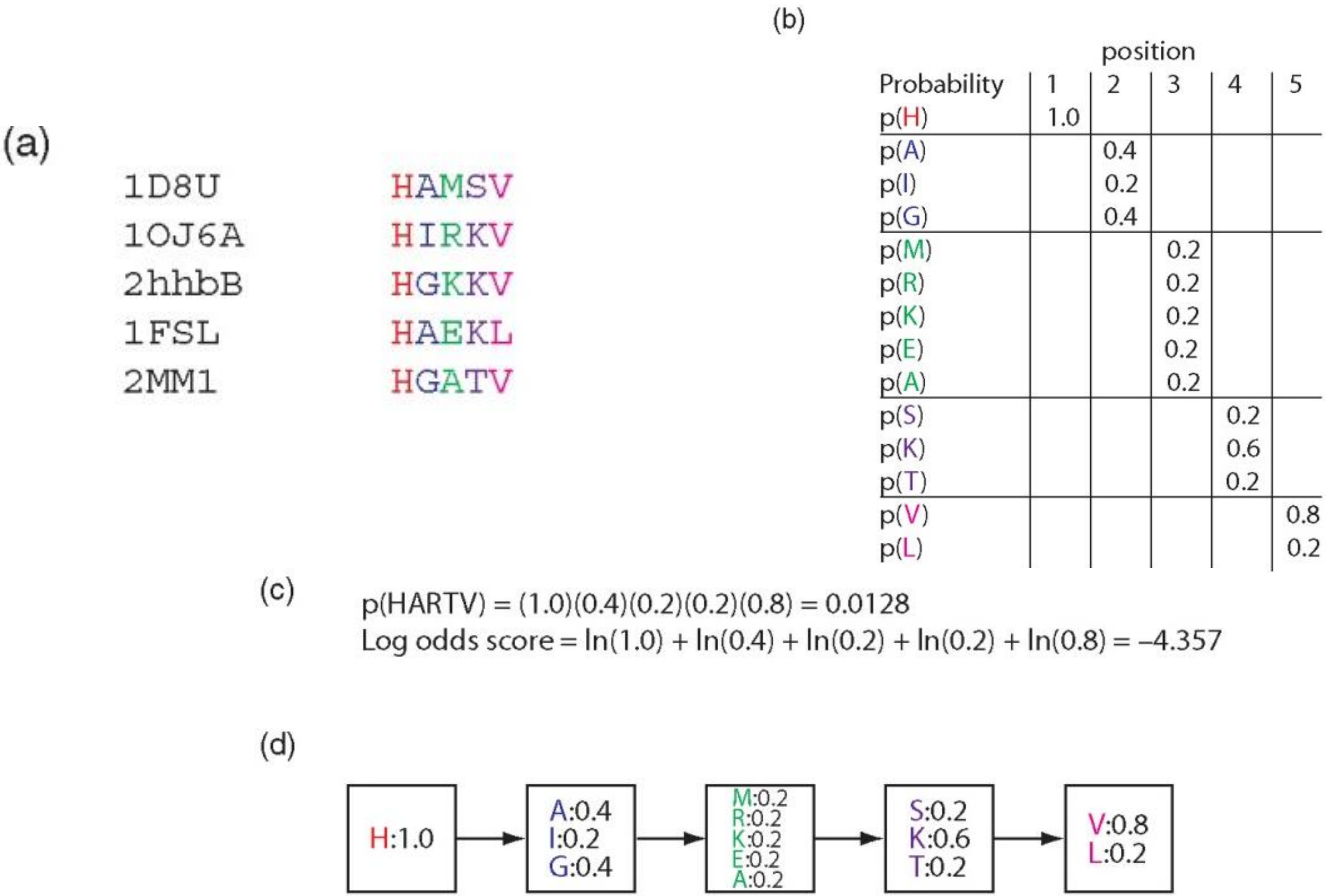
We can further describe the probability of occurrence of a protein sequence we have not encountered (e.g. HARTV)

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) =



We can further describe the probability of occurrence of a protein sequence we have not encountered (e.g. HARTV).



(a)

1D8U	HAMSV
1OJ6A	HIRKV
2hhbB	HGKKV
1FSL	HAEKL
2MM1	HGATV

(b)

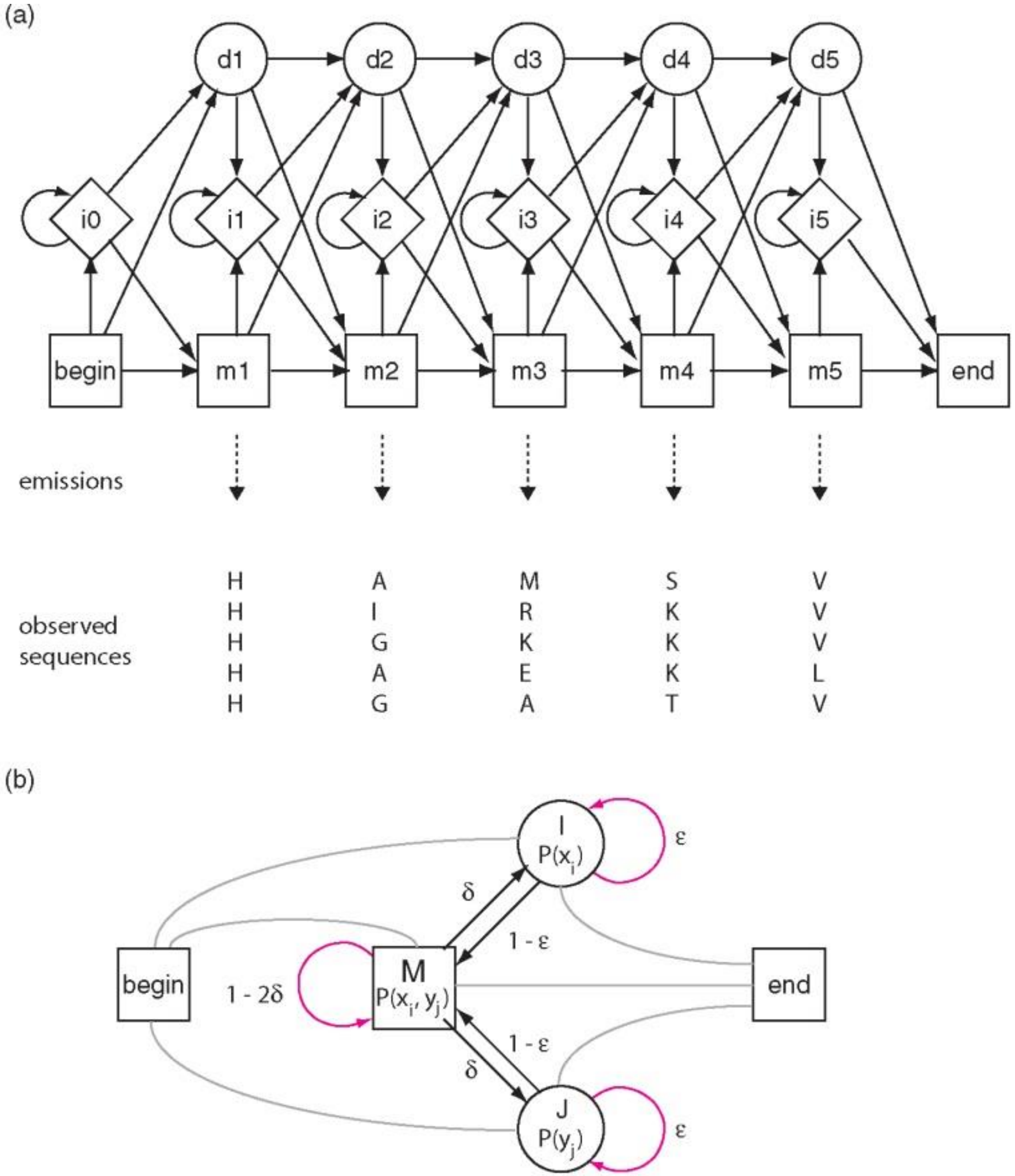
	position				
Probability	1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

(c)

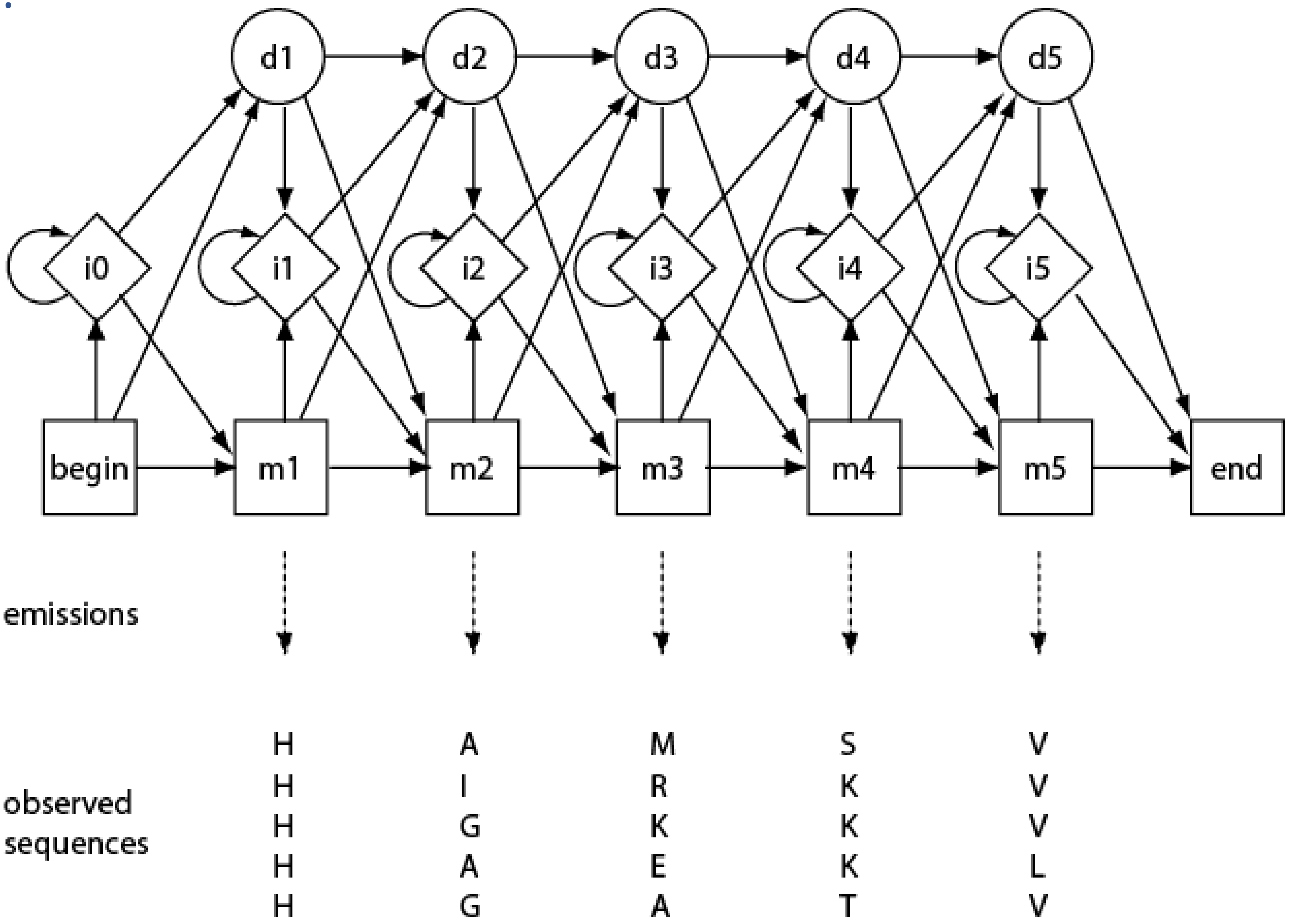
p(HARTV) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128

Log odds score = ln(1.0) + ln(0.4) + ln(0.2) + ln(0.2) + ln(0.8) = –4.357

(d)

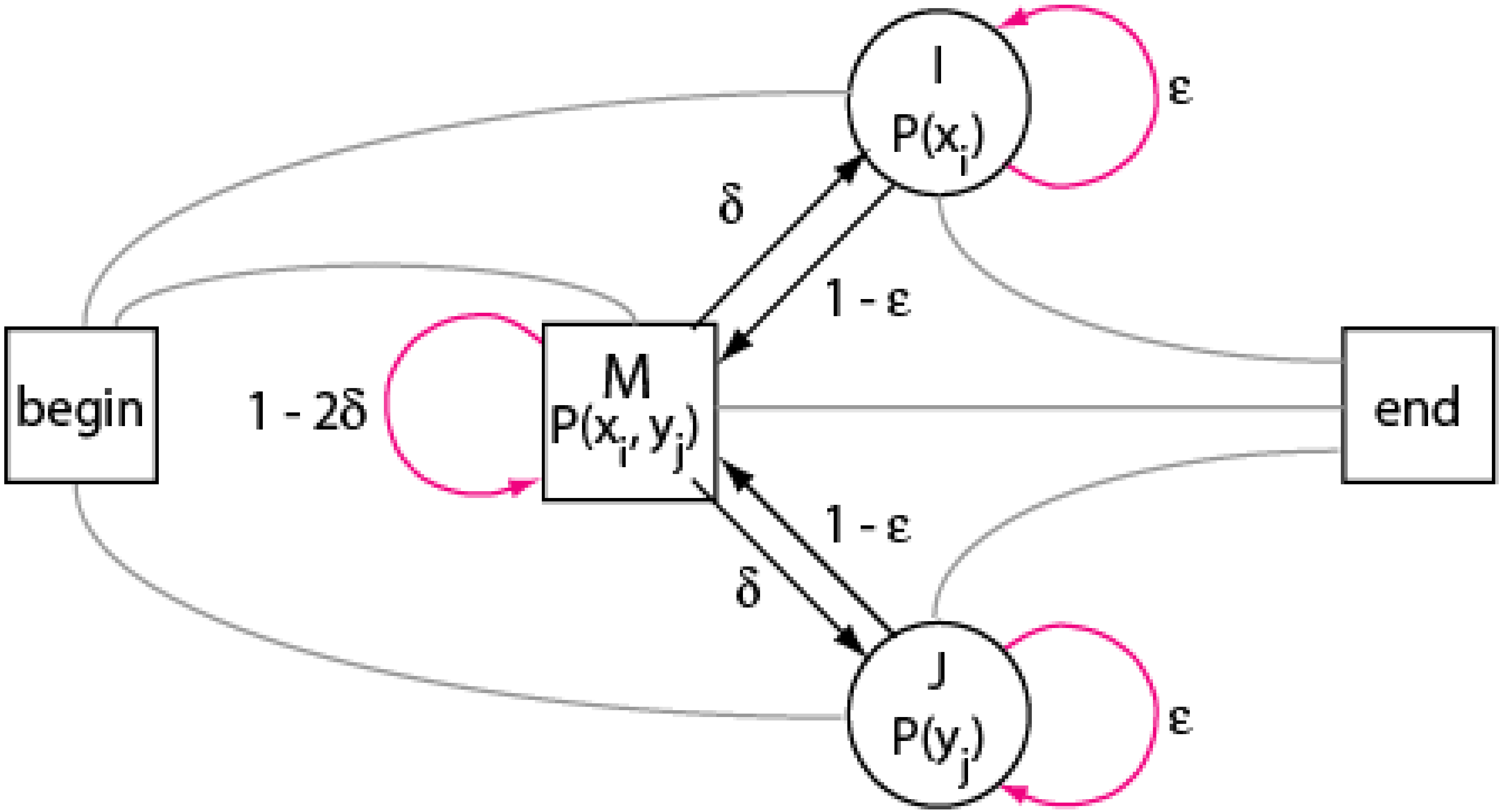

(a)
d1
d2
d3
d4
d5
i0
i1
i2
i3
i4
i5
begin
m1
m2
m3
m4
m5
end
emissions
observed sequences
H A M S V
H I R K V
H G K K V
H A E K L
H G A T V
(b)
I
P(x_i)
ε
δ
1 - ε
begin
1 - 2δ
M
P(x_i, y_j)
end
1 - ε
δ
J
P(y_j)
ε

A hidden Markov model (HMM) includes beginning and end states, insertion and deletion states, and probabilities that explain the observed sequences.



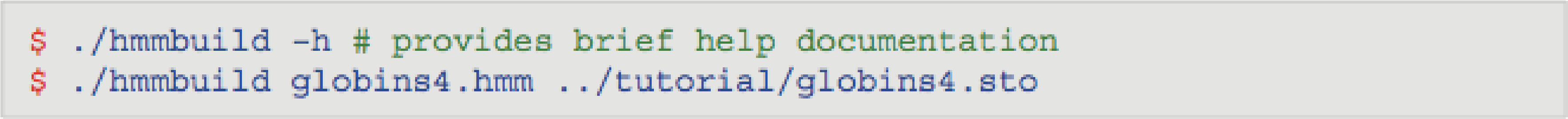
H A M S V
H I R K V
H G K K V
H A E K L
H G A T V

A pairwise HMM describes how two sequences are aligned



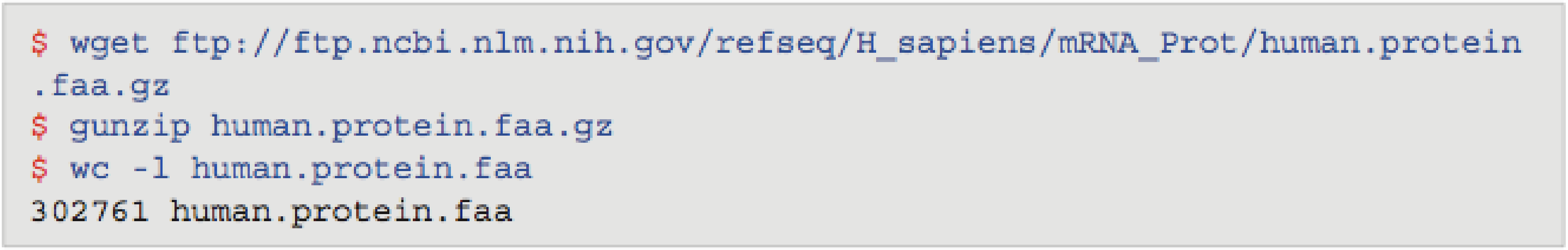
HMMER software: build profiles, complement BLAST

Build a profile HMM (input is a multiple sequence alignment)



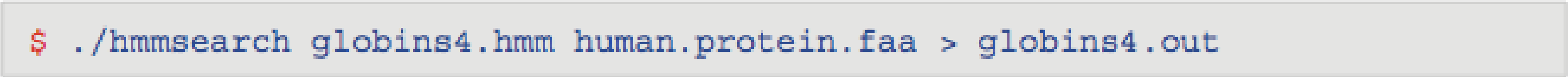
```
$ ./hmmbuild -h # provides brief help documentation
$ ./hmmbuild globins4.hmm ../tutorial/globins4.sto
```

Download a database to search (e.g. human RefSeq proteins)



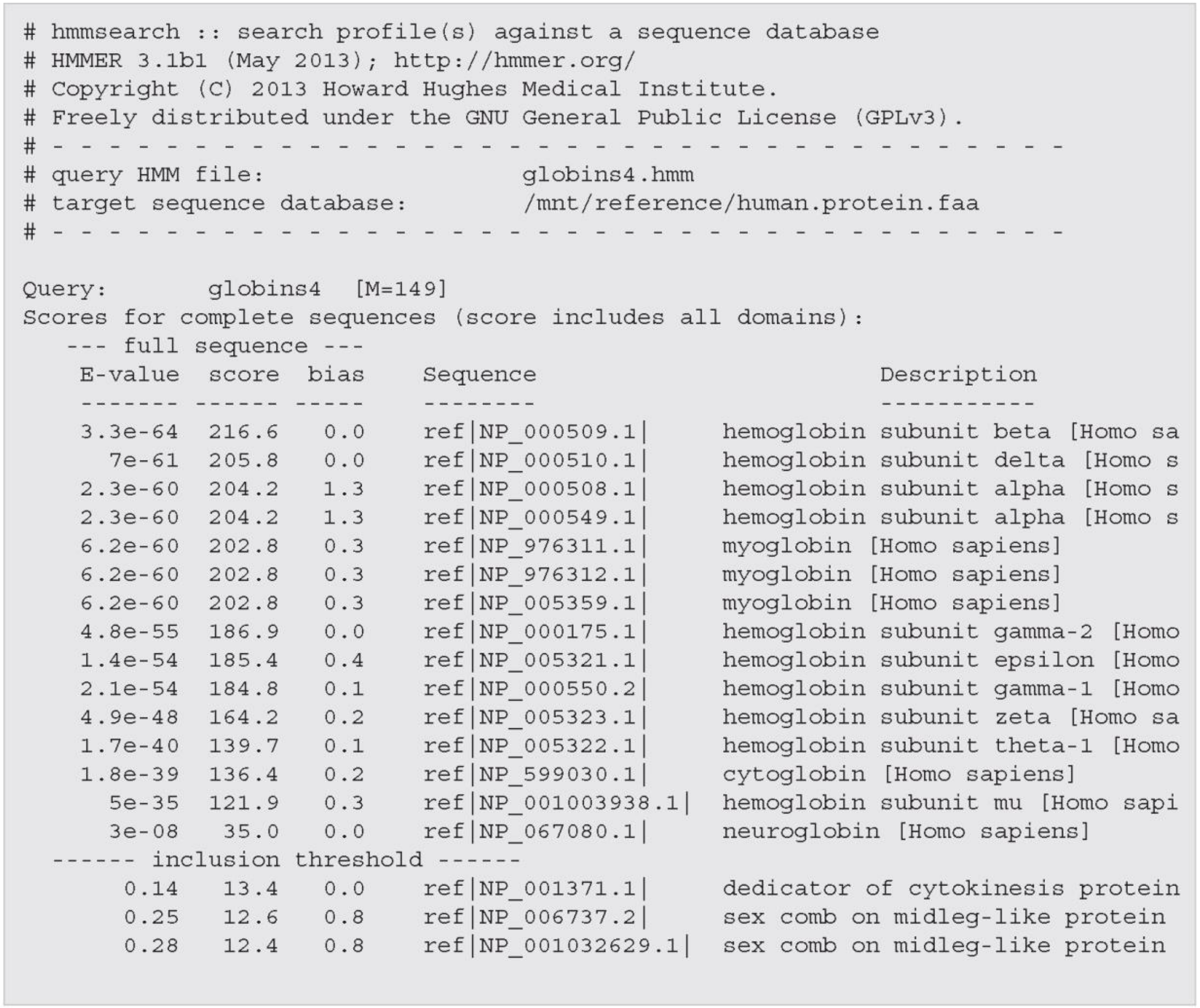
```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein
.faa.gz
$ gunzip human.protein.faa.gz
$ wc -l human.protein.faa
302761 human.protein.faa
```

Search an HMM against a database



```
$ ./hmmsearch globins4.hmm human.protein.faa > globins4.out
```

Use HMMER to build a profile HMM then search a database



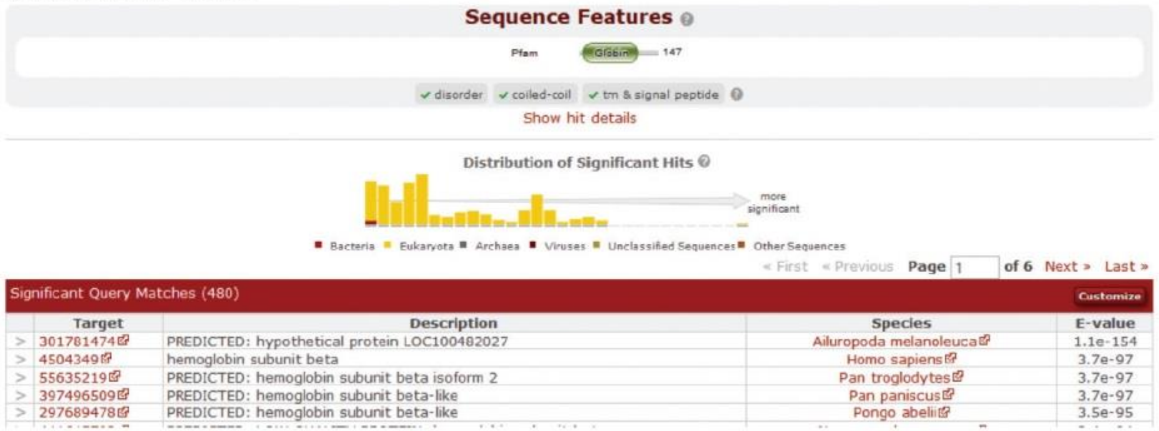
```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  globins4.hmm
# target sequence database:        /mnt/reference/human.protein.faa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---
    E-value  score  bias    Sequence                        Description
    ------- ------ -----    --------                        -----------
    3.3e-64  216.6   0.0    ref|NP_000509.1|     hemoglobin subunit beta [Homo sa
      7e-61  205.8   0.0    ref|NP_000510.1|     hemoglobin subunit delta [Homo s
    2.3e-60  204.2   1.3    ref|NP_000508.1|     hemoglobin subunit alpha [Homo s
    2.3e-60  204.2   1.3    ref|NP_000549.1|     hemoglobin subunit alpha [Homo s
    6.2e-60  202.8   0.3    ref|NP_976311.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8   0.3    ref|NP_976312.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8   0.3    ref|NP_005359.1|     myoglobin [Homo sapiens]
    4.8e-55  186.9   0.0    ref|NP_000175.1|     hemoglobin subunit gamma-2 [Homo
    1.4e-54  185.4   0.4    ref|NP_005321.1|     hemoglobin subunit epsilon [Homo
    2.1e-54  184.8   0.1    ref|NP_000550.2|     hemoglobin subunit gamma-1 [Homo
    4.9e-48  164.2   0.2    ref|NP_005323.1|     hemoglobin subunit zeta [Homo sa
    1.7e-40  139.7   0.1    ref|NP_005322.1|     hemoglobin subunit theta-1 [Homo
    1.8e-39  136.4   0.2    ref|NP_599030.1|     cytoglobin [Homo sapiens]
      5e-35  121.9   0.3    ref|NP_001003938.1|  hemoglobin subunit mu [Homo sapi
      3e-08   35.0   0.0    ref|NP_067080.1|     neuroglobin [Homo sapiens]
  ------ inclusion threshold ------
       0.14   13.4   0.0    ref|NP_001371.1|     dedicator of cytokinesis protein
       0.25   12.6   0.8    ref|NP_006737.2|     sex comb on midleg-like protein
       0.28   12.4   0.8    ref|NP_001032629.1|  sex comb on midleg-like protein
```

HMMER output includes scores, E values

HMMER is available online

(a) HMMER web output

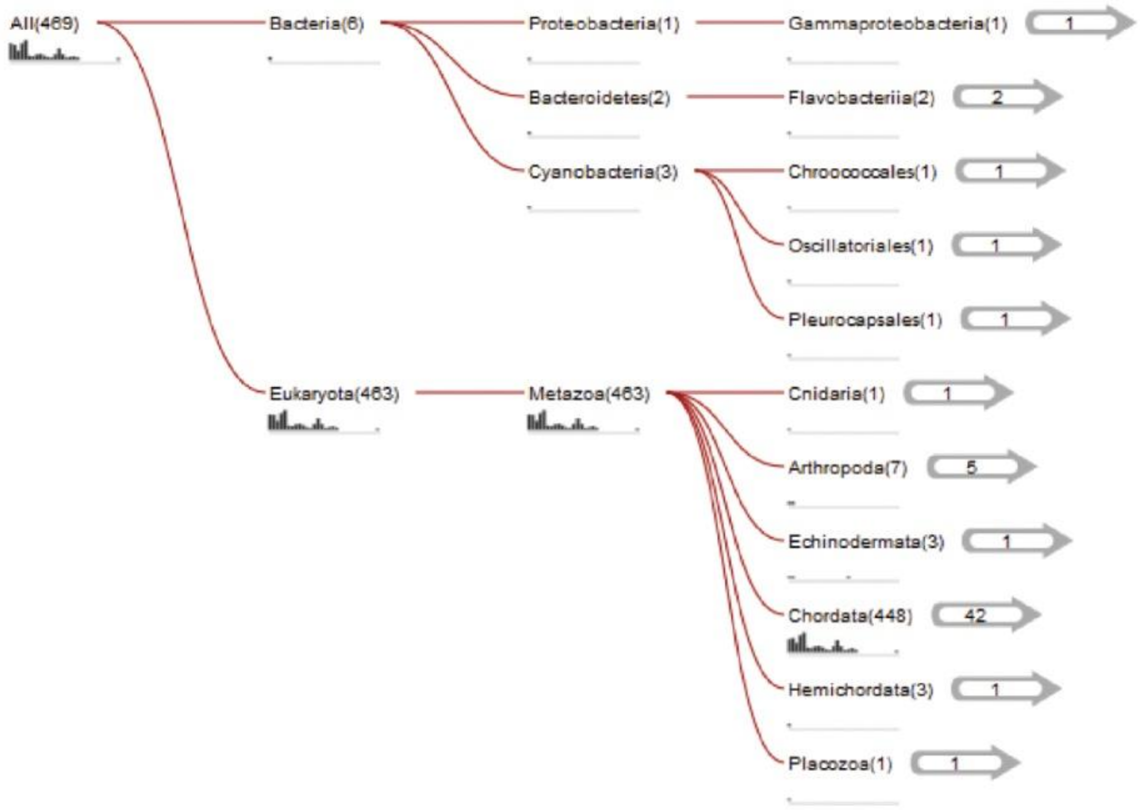


Sequence Features
Pfam
Globin
147
✓ disorder
✓ coiled-coil
✓ tm & signal peptide
Show hit details
Distribution of Significant Hits
more significant
■ Bacteria ■ Eukaryota ■ Archaea ■ Viruses ■ Unclassified Sequences ■ Other Sequences
« First « Previous Page 1 of 6 Next » Last »
Customize

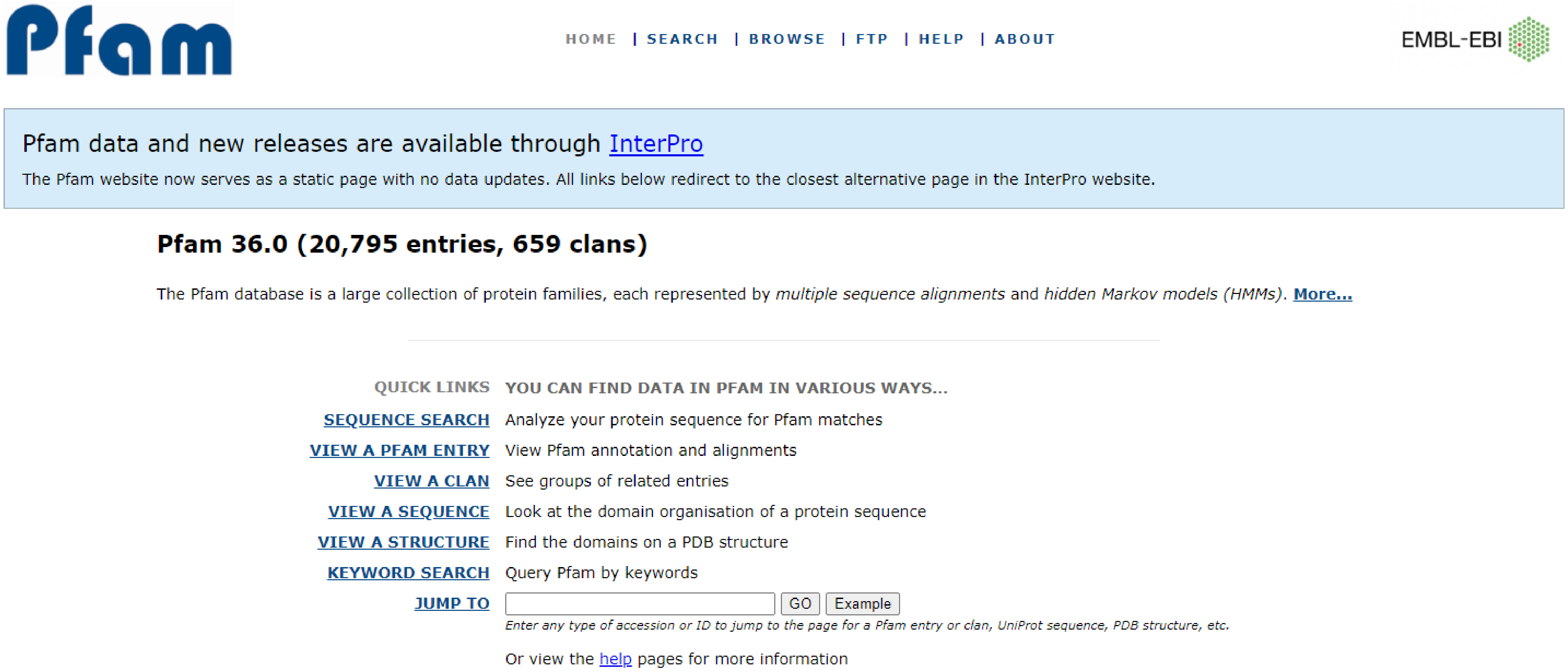
Significant Query Matches (480)

	Target	Description	Species	E-value
>	301781474	PREDICTED: hypothetical protein LOC100482027	Ailuropoda melanoleuca	1.1e-154
>	4504349	hemoglobin subunit beta	Homo sapiens	3.7e-97
>	55635219	PREDICTED: hemoglobin subunit beta isoform 2	Pan troglodytes	3.7e-97
>	397496509	PREDICTED: hemoglobin subunit beta-like	Pan paniscus	3.7e-97
>	297689478	PREDICTED: hemoglobin subunit beta-like	Pongo abelii	3.5e-95

(b) HMMER phylogenetic output



PFAM is a database of HMMs and an essential resource for protein families http://pfam.xfam.org/



Pfam data and new releases are available through InterPro

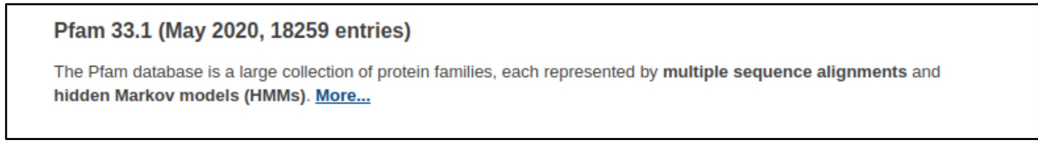
The Pfam website now serves as a static page with no data updates. All links below redirect to the closest alternative page in the InterPro website.

Pfam 36.0 (20,795 entries, 659 clans)

The Pfam database is a large collection of protein families, each represented by *multiple sequence alignments* and *hidden Markov models (HMMs)*. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	*Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.*

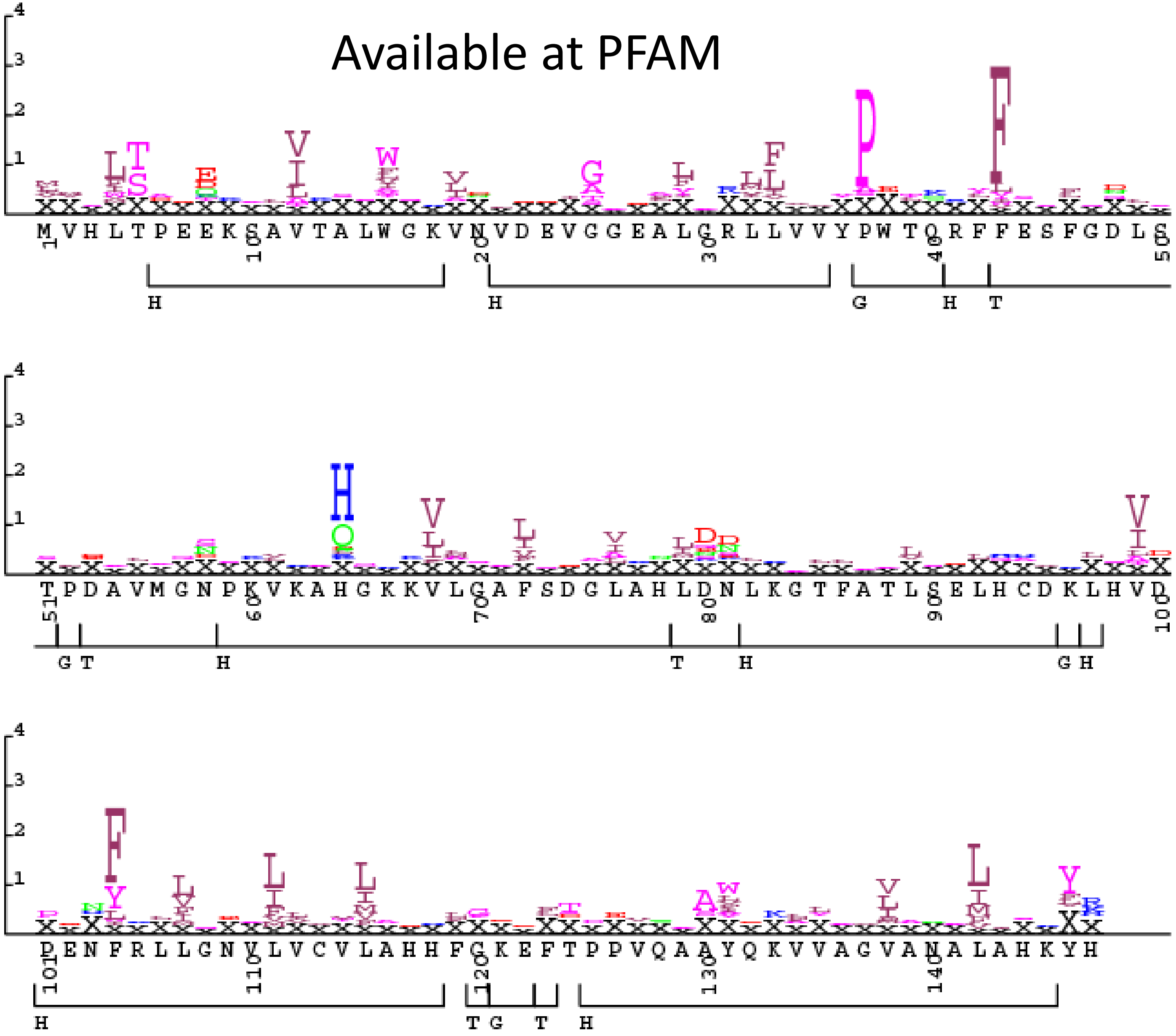
Or view the help pages for more information



Pfam 33.1 (May 2020, 18259 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

HMM logos graphically depict the likelihood of observed amino acids



Available at PFAM

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~~

PatternHunter uses long seeds with mismatches to improve sensitivity

(a)

11111111111
ATGGTGCATCT (example of a seed) (extended)
ATTGTGCATCT (example of a mismatch) (not extended)

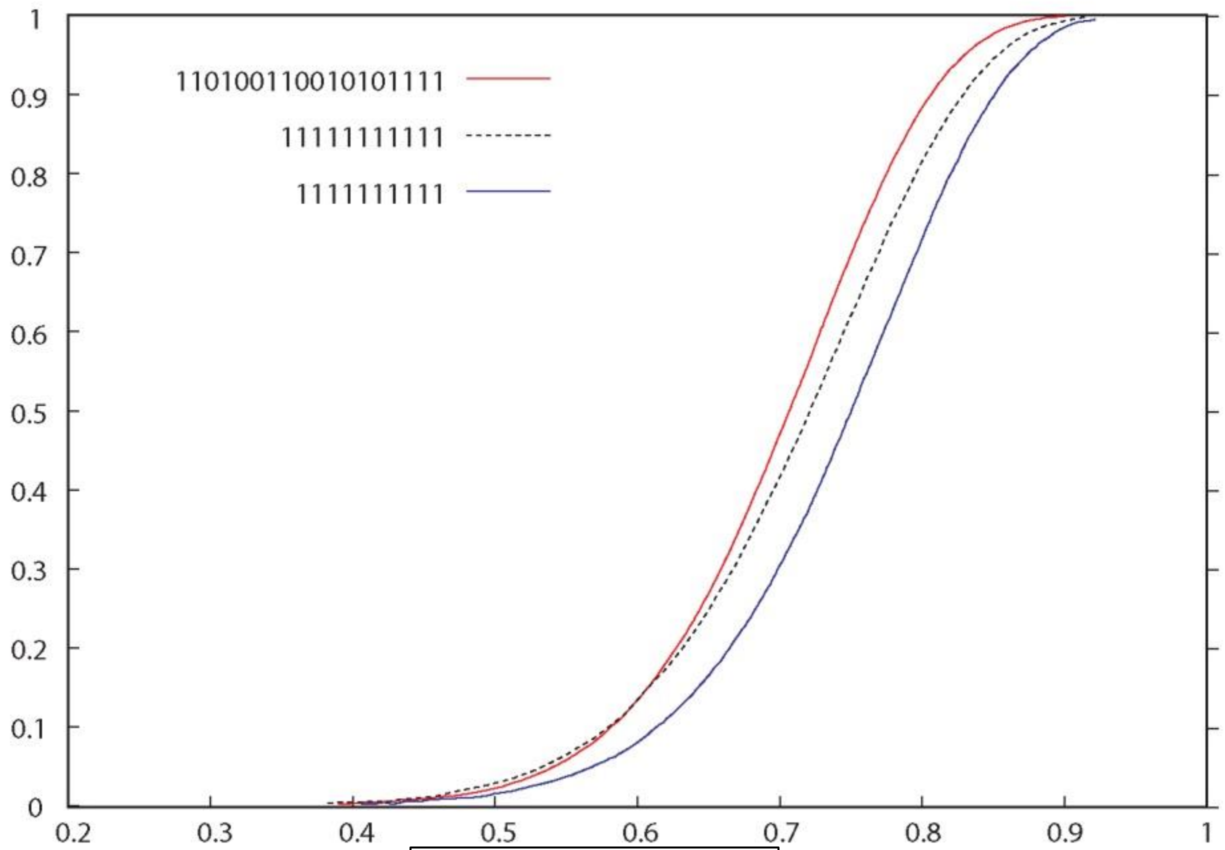
11111111111

(b)

110100110010101111
ATGGTGCATCTGACTCCT (example of a seed) (extended)
ATTGTGCATCTGACTCCT (example of an acceptable match) (extended)

110100110010101111

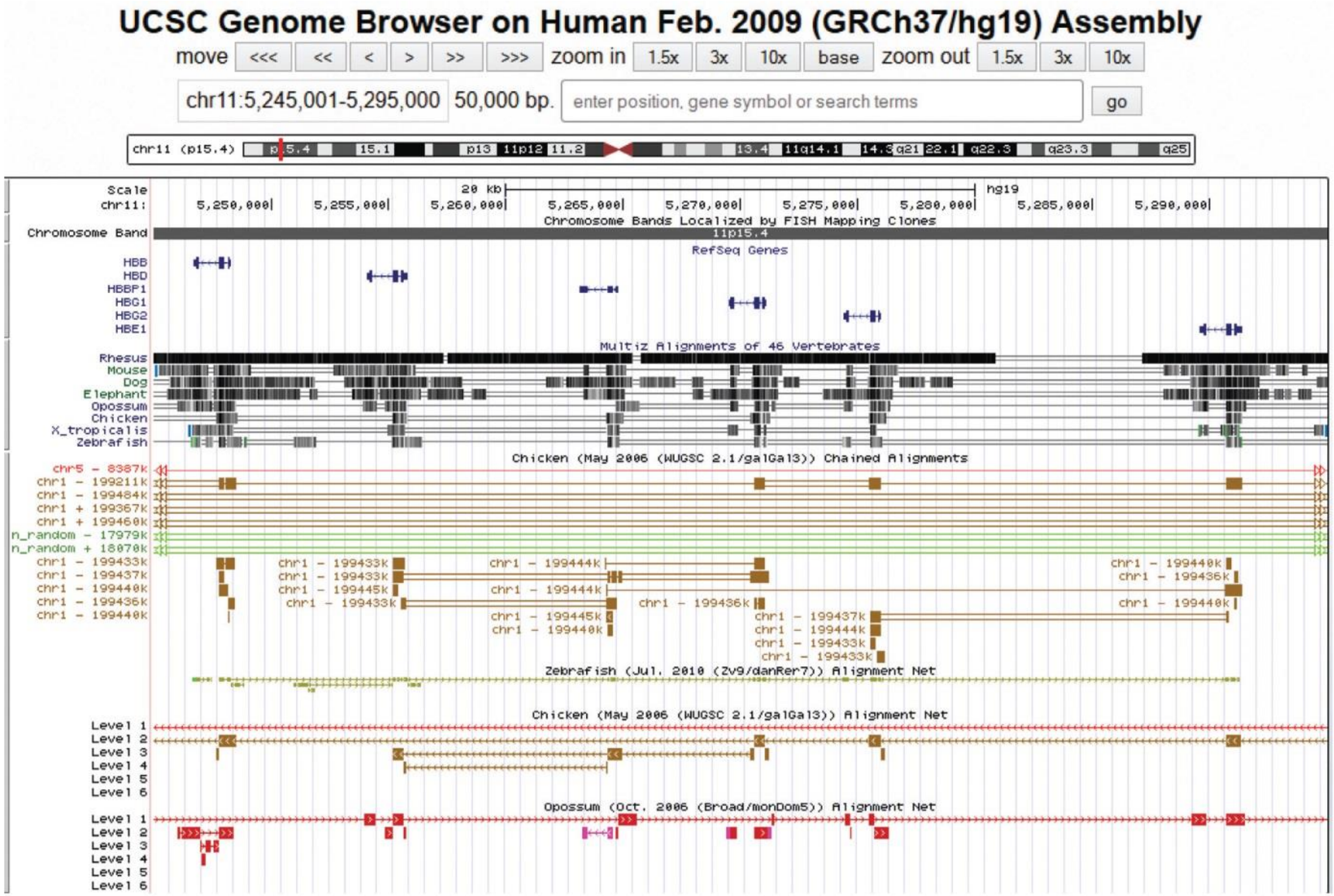
(c)



sensitivity

similarity

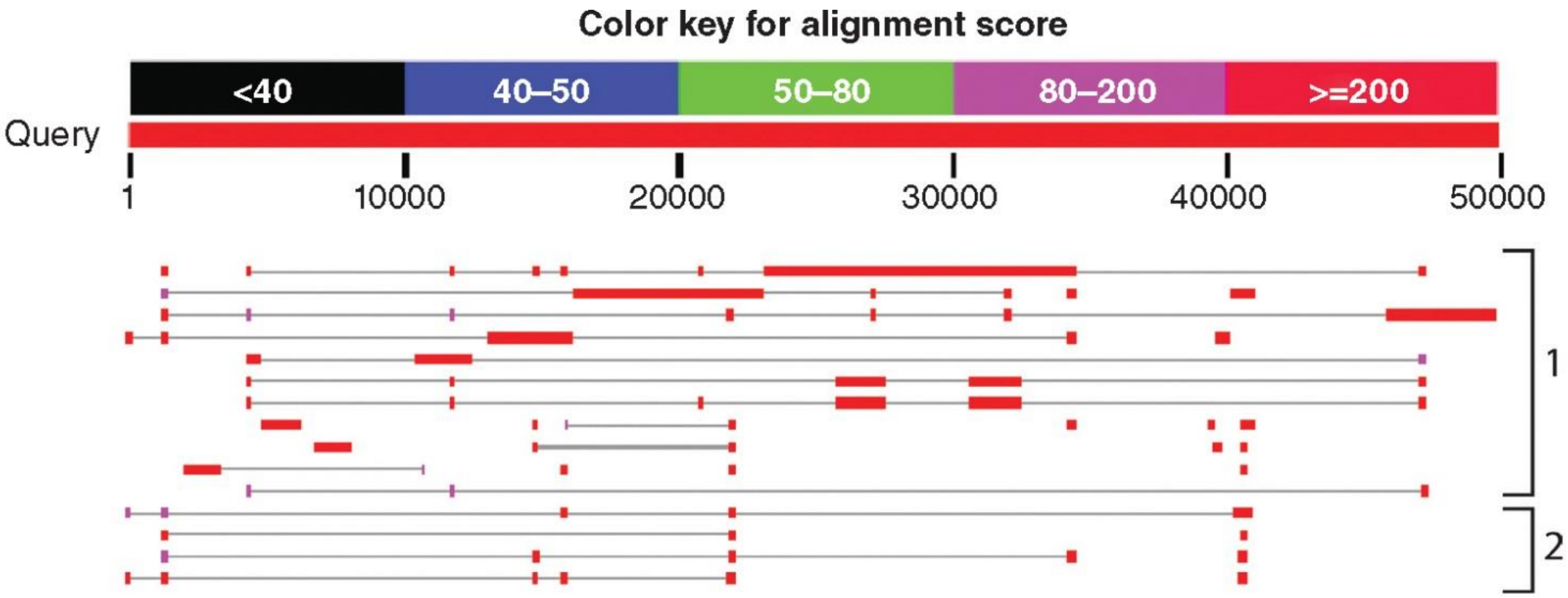
BLASTZ alignments at UCSC (replaced by LASTZ)

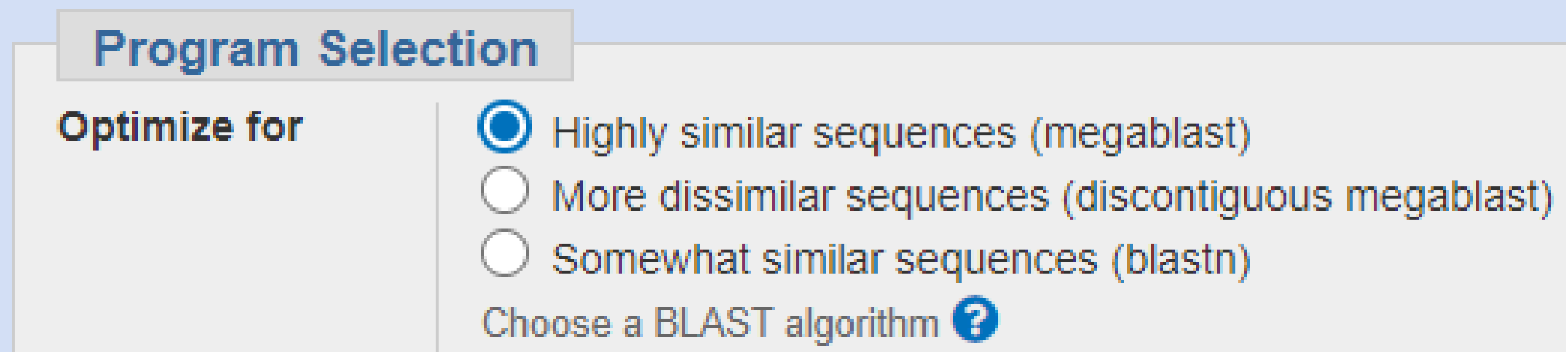


50 kilobases at the beta globin locus are displayed, including BLASTZ alignments.

MegaBLAST: extremely fast searches with large seeds

- very fast
- uses very large word sizes (e.g. w=28, up to w=256)
- use it to align long, closely related sequences
- choose discontiguous megablast for cross-species comparisons (tolerates mismatches)





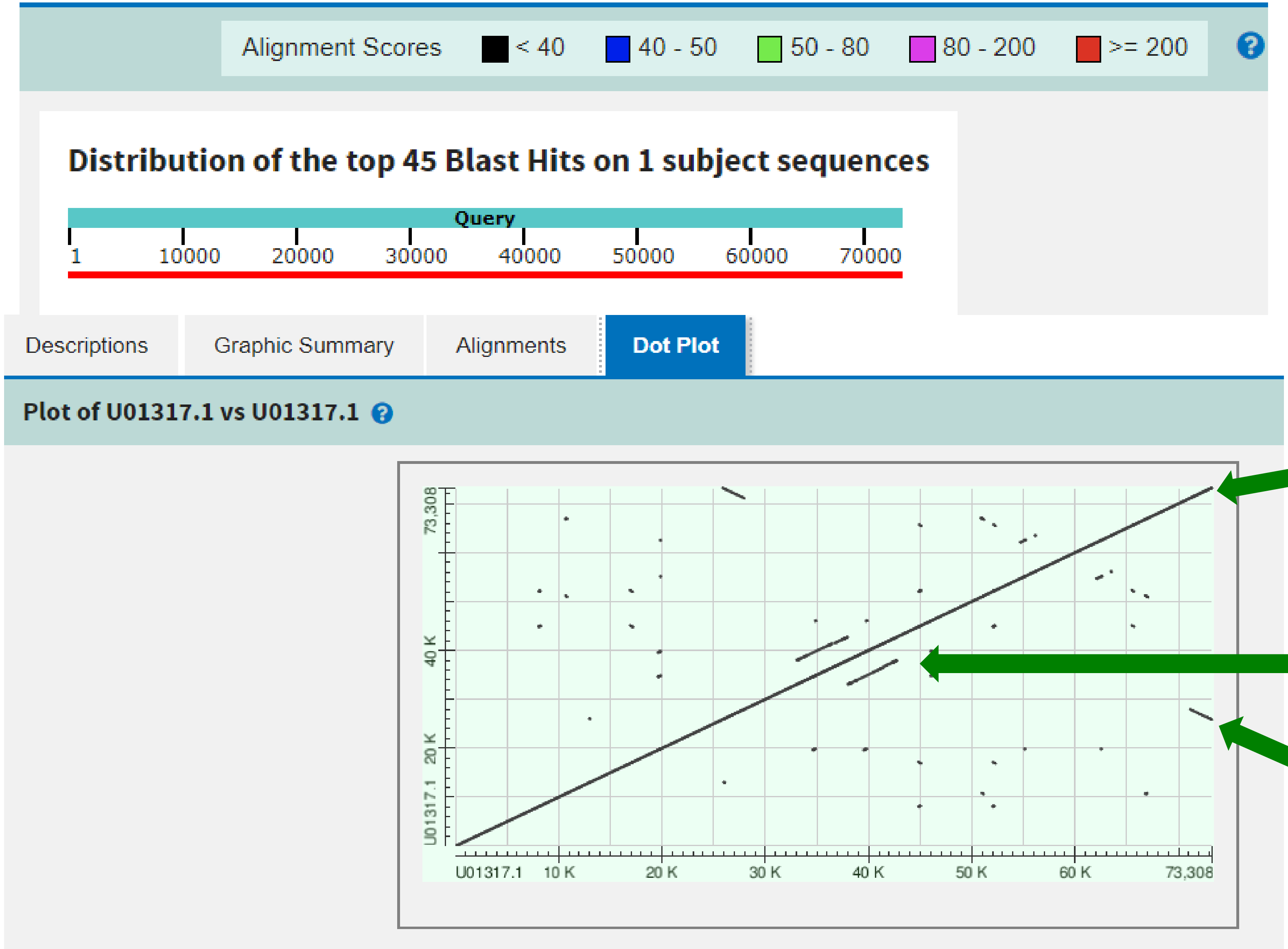
Program Selection

Optimize for

● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

MegaBLAST output



Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200

Distribution of the top 45 Blast Hits on 1 subject sequences

Descriptions | Graphic Summary | Alignments | **Dot Plot**

Plot of U01317.1 vs U01317.1

60 kb of the human beta globin locus versus itself!

main diagonal reflects exact matches

off diagonal reflects duplications

Flipped, off diagonal reflects inversions

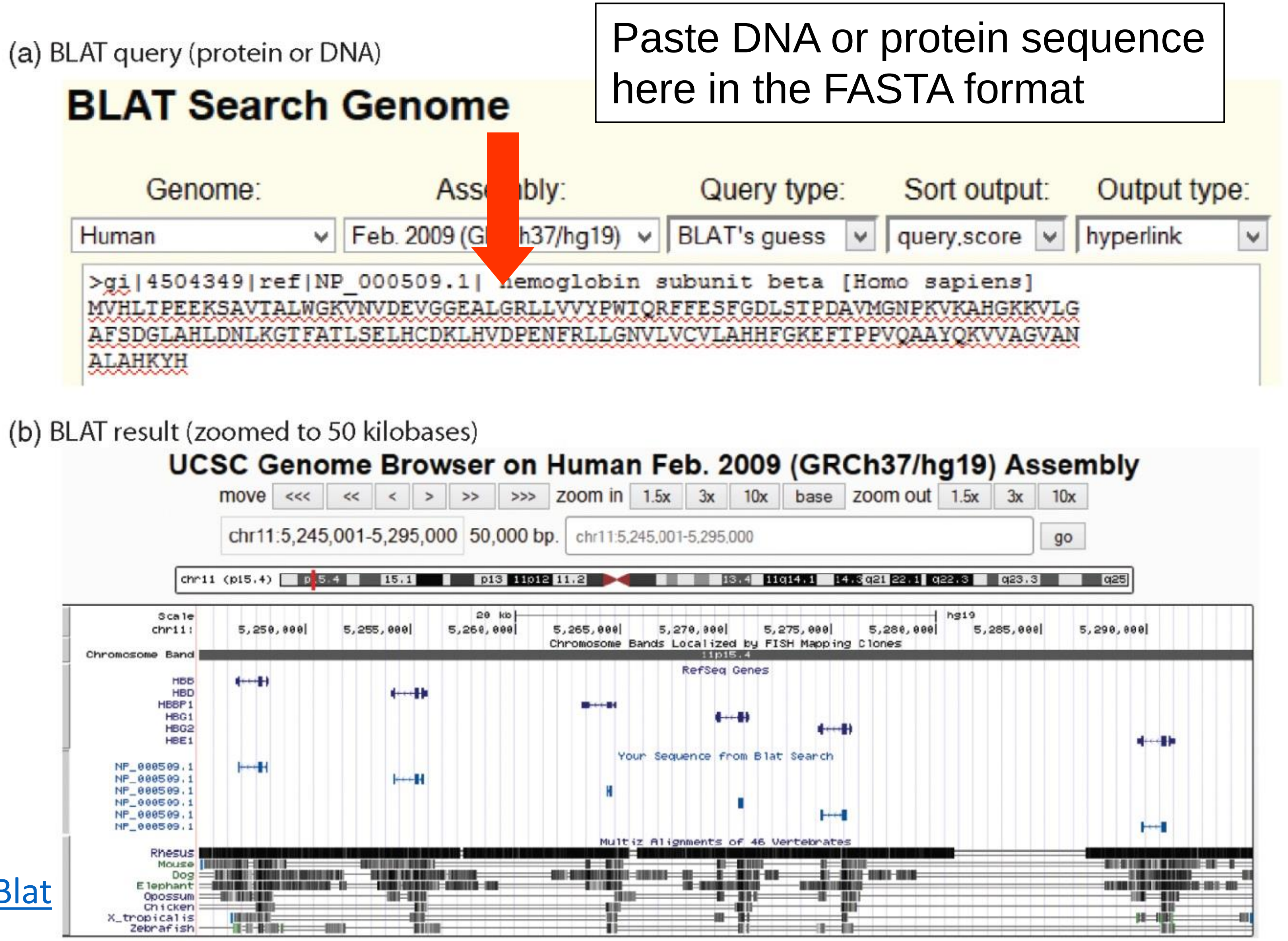
BLAT indexes a whole genomic database rather than a query

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.
BLAT on proteins finds sequences of 80% and greater similarity of length 20+ amino acids. In practice DNA BLAT works well on primates, and protein blat on terrestrial vertebrates.

~BLAT website

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC187518/

BLAT indexes a whole genomic database rather than a query



(a) BLAT query (protein or DNA)

Paste DNA or protein sequence here in the FASTA format

BLAT Search Genome

Genome: Human
Assembly: Feb. 2009 (GRCh37/hg19)
Query type: BLAT's guess
Sort output: query,score
Output type: hyperlink

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

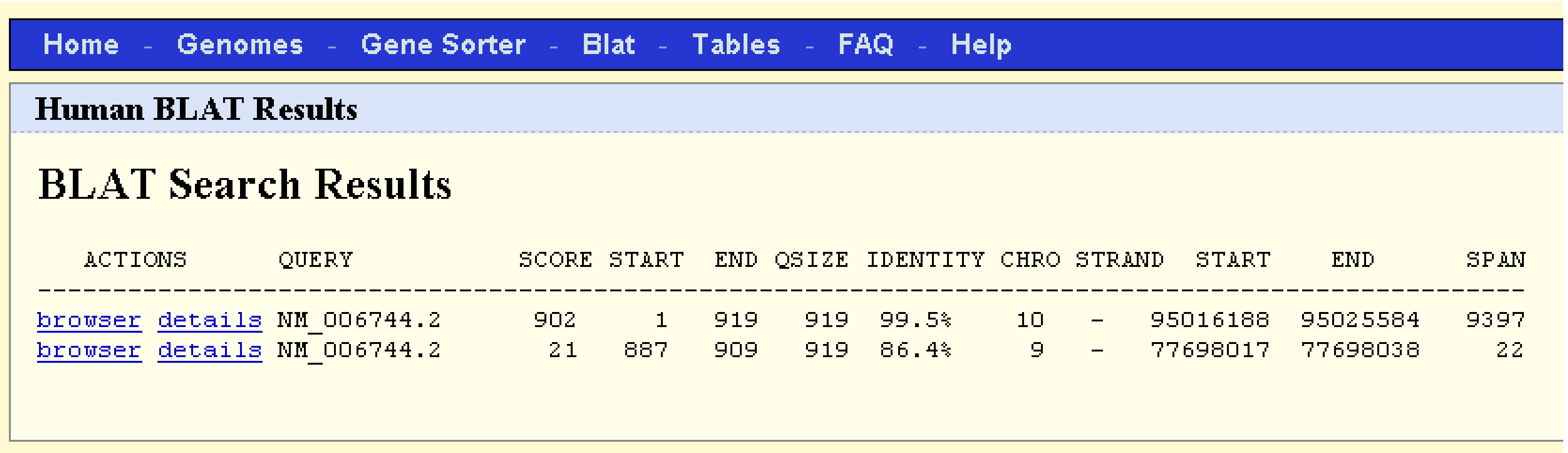
(b) BLAT result (zoomed to 50 kilobases)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:5,245,001-5,295,000 50,000 bp.

https://genome.ucsc.edu/cgi-bin/hgBlat

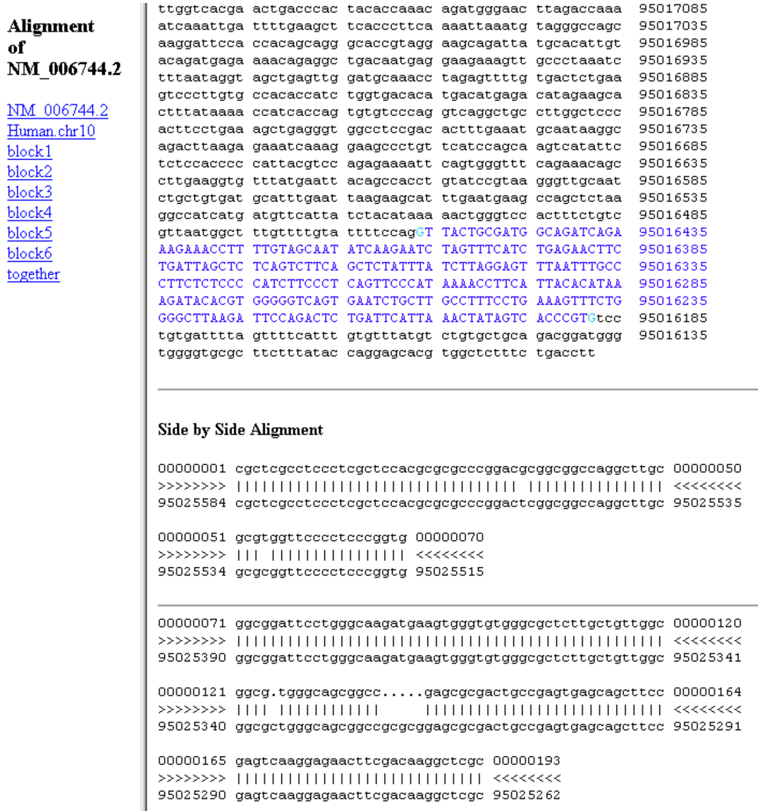
BLAT output includes browser and other formats



Human BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START     END      SPAN
-----------------------------------------------------------------------------------------------------
browser details NM_006744.2       902     1   919   919  99.5%   10   -    95016188 95025584   9397
browser details NM_006744.2        21   887   909   919  86.4%    9   -    77698017 77698038     22
```



Alignment of NM_006744.2

NM_006744.2
Human.chr10
block1
block2
block3
block4
block5
block6
together

```
ttggtcacga actgacccac tacaccaaac agatgggaac ttagaccaaa  95017085
atcaaattga ttttgaagct tcacccttca aaattaaatg tagggccagc  95017035
aaggattcca ccacagcagg gcaccgtagg aagcagatta tgcacattgt  95016985
acagatgaga aaacagaggc tgacaatgag gaagaaagtt gccctaaatc  95016935
tttaataggt agctgagttg gatgcaaacc tagagttttg tgactctgaa  95016885
gtcccttgtg ccacaccatc tggtgacaca tgacatgaga catagaagca  95016835
ctttataaaa ccatcaccag tgtgtcccag gtcaggctgc cttggctccc  95016785
acttcctgaa agctgagggt ggcctccgac actttgaaat gcaataaggc  95016735
agacttaaga gaaatcaaag gaagccctgt tcatccagca agtcatattc  95016685
tctccacccc cattacgtcc agagaaaatt cagtgggttt cagaaacagc  95016635
cttgaaggtg tttatgaatt acagccacct gtatccgtaa gggttgcaat  95016585
ctgctgtgat gcatttgaat taagaagcat ttgaatgaag ccagctctaa  95016535
ggccatcatg atgttcatta tctacataaa aactgggtcc actttctgtc  95016485
gttaatggct ttgttttgta ttttccagGT TACTGCGATG GCAGATCAGA  95016435
AAGAAACCTT TTGTAGCAAT ATCAAGAATC TAGTTTCATC TGAGAACTTC  95016385
TGATTAGCTC TCAGTCTTCA GCTCTATTTA TCTTAGGAGT TTAATTTGCC  95016335
CTTCTCTCCC CATCTTCCCT CAGTTCCCAT AAAACCTTCA TTACACATAA  95016285
AGATACACGT GGGGGTCAGT GAATCTGCTT GCCTTTCCTG AAAGTTTCTG  95016235
GGGCTTAAGA TTCCAGACTC TGATTCATTA AACTATAGTC ACCCGTGtcc  95016185
tgtgatttta gttttcattt gtgtttatgt ctgtgctgca gacggatggg  95016135
tggggtgcgc ttctttatac caggagcacg tggctctttc tgacctt
```

Side by Side Alignment

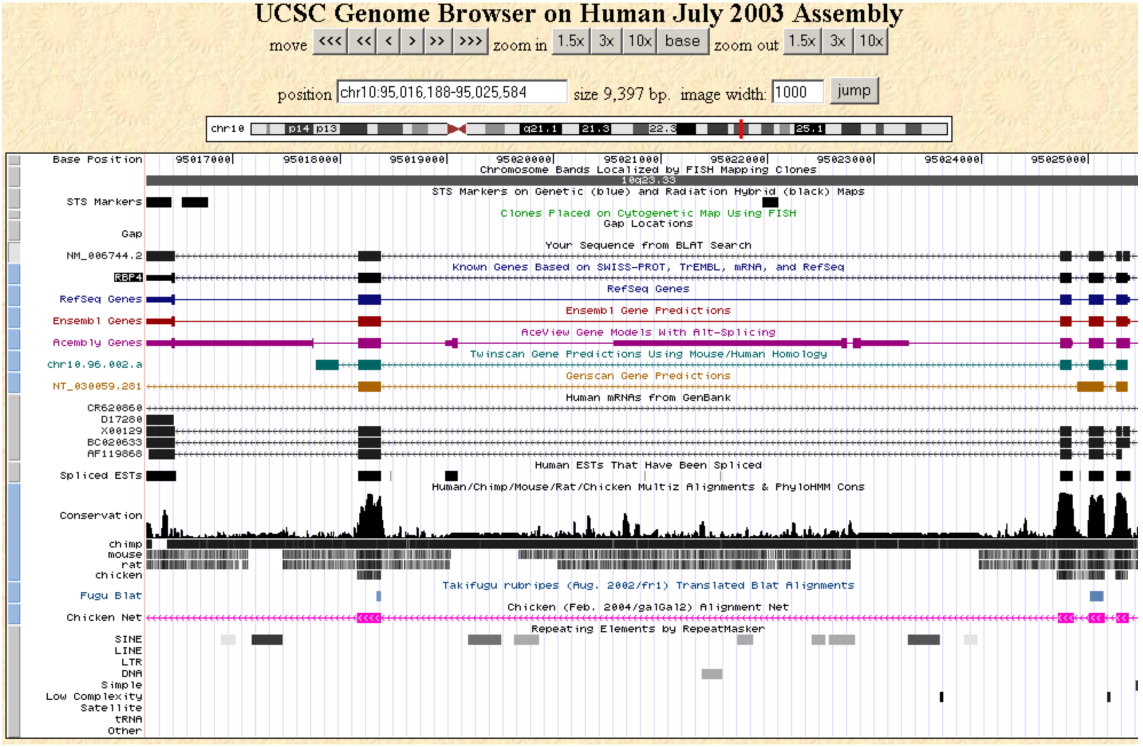
```
00000001 cgctcgcctccctcgctccacgcgcgcccggacgcggcggccaggcttgc 00000050
>>>>>>>> ||||||||||||||||||||||||||||||||| |||||||||||||||| <<<<<<<<
95025584 cgctcgcctccctcgctccacgcgcgcccggactcggcggccaggcttgc 95025535

00000051 gcgtggttcccctcccggtg 00000070
>>>>>>>> ||| |||||||||||||||| <<<<<<<<
95025534 gcgcggttcccctcccggtg 95025515
```

```
00000071 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 00000120
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<
95025390 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 95025341

00000121 ggcg.tgggcagcggcc.....gagcgcgactgccgagtgagcagcttcc 00000164
>>>>>>>> |||| ||||||||||||     |||||||||||||||||||||||||||| <<<<<<<<
95025340 ggcgctgggcagcggccgcgcggagcgcgactgccgagtgagcagcttcc 95025291

00000165 gagtcaaggagaacttcgacaaggctcgc 00000193
>>>>>>>> ||||||||||||||||||||||||||||| <<<<<<<<
95025290 gagtcaaggagaacttcgacaaggctcgc 95025262
```

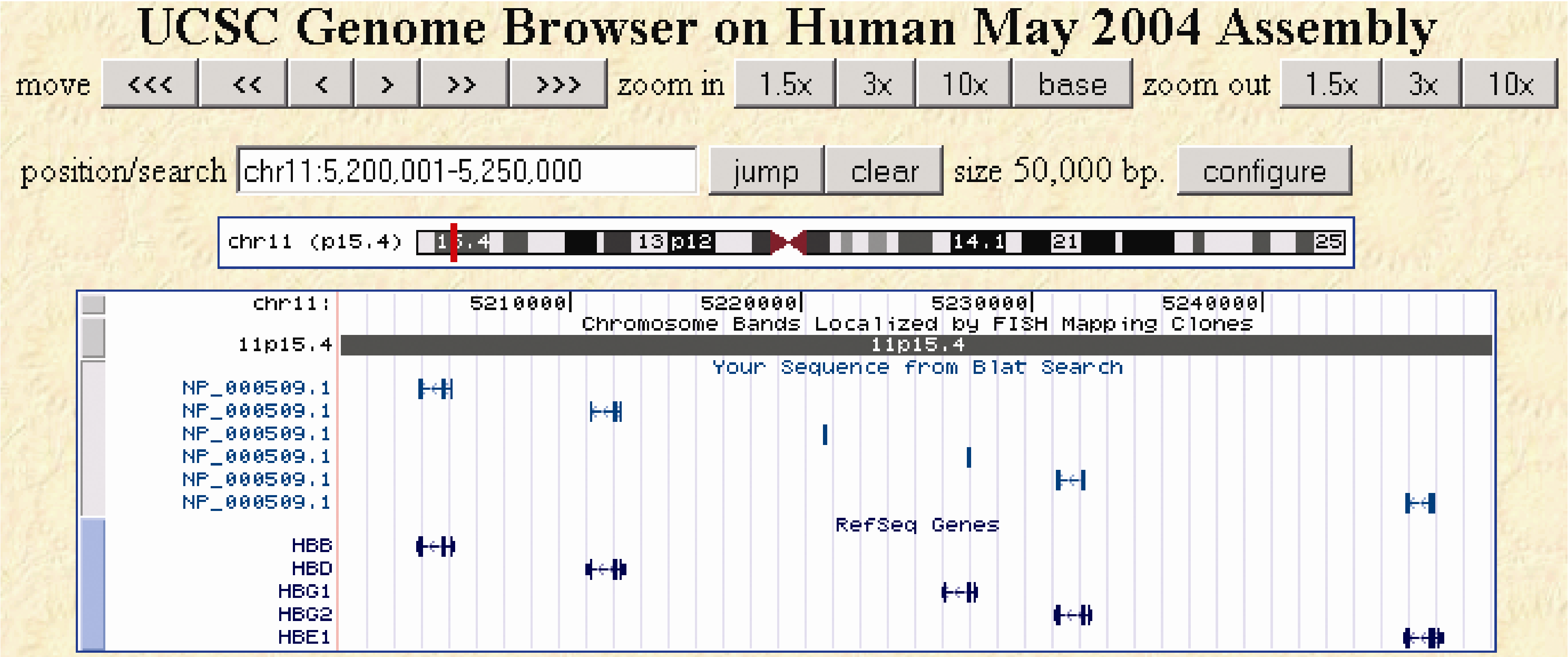


UCSC Genome Browser on Human July 2003 Assembly

position chr10:95,016,188-95,025,584 size 9,397 bp. image width: 1000

move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
jump
chr10
p14
p13
q21.1
21.3
22.3
25.1
Base Position
95017000
95018000
95019000
95020000
95021000
95022000
95023000
95024000
95025000
Chromosome Bands Localized by FISH Mapping Clones
10q23.33
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Clones Placed on Cytogenetic Map Using FISH
Gap Locations
Gap
Your Sequence from BLAT Search
NM_006744.2
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
RBP4
RefSeq Genes
RefSeq Genes
Ensembl Gene Predictions
Ensembl Genes
AceView Gene Models With Alt-Splicing
Acembly Genes
Twinscan Gene Predictions Using Mouse/Human Homology
chr10.96.002.a
Genscan Gene Predictions
NT_030059.281
Human mRNAs from GenBank
CR620860
D17280
X00129
BC020633
AF119868
Human ESTs That Have Been Spliced
Spliced ESTs
Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
chimp
mouse
rat
chicken
Takifugu rubripes (Aug. 2002/fr1) Translated Blat Alignments
Fugu Blat
Chicken (Feb. 2004/galGal2) Alignment Net
Chicken Net
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
tRNA
Other

BLAT output includes browser and other formats

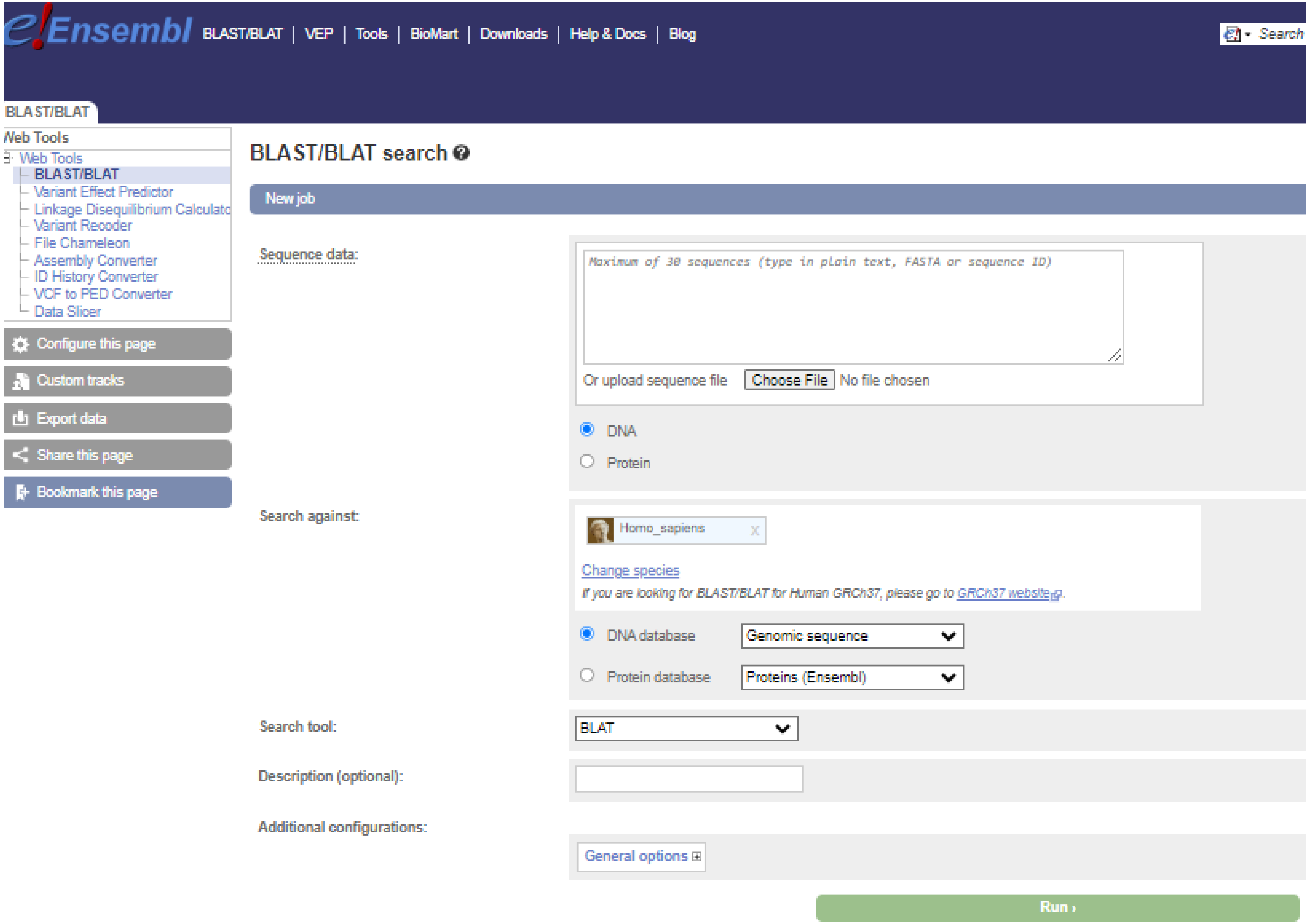


UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000

size 50,000 bp.

This example shows a BLAT query of beta globin resulting in a series of matches to homologous, neighboring globins.



BLAST/BLAT search

Sequence data: Maximum of 30 sequences (type in plain text, FASTA or sequence ID)

Or upload sequence file Choose File No file chosen

● DNA

○ Protein

Search against: Homo_sapiens x

Change species

If you are looking for BLAST/BLAT for Human GRCh37, please go to GRCh37 website.

● DNA database Genomic sequence

○ Protein database Proteins (Ensembl)

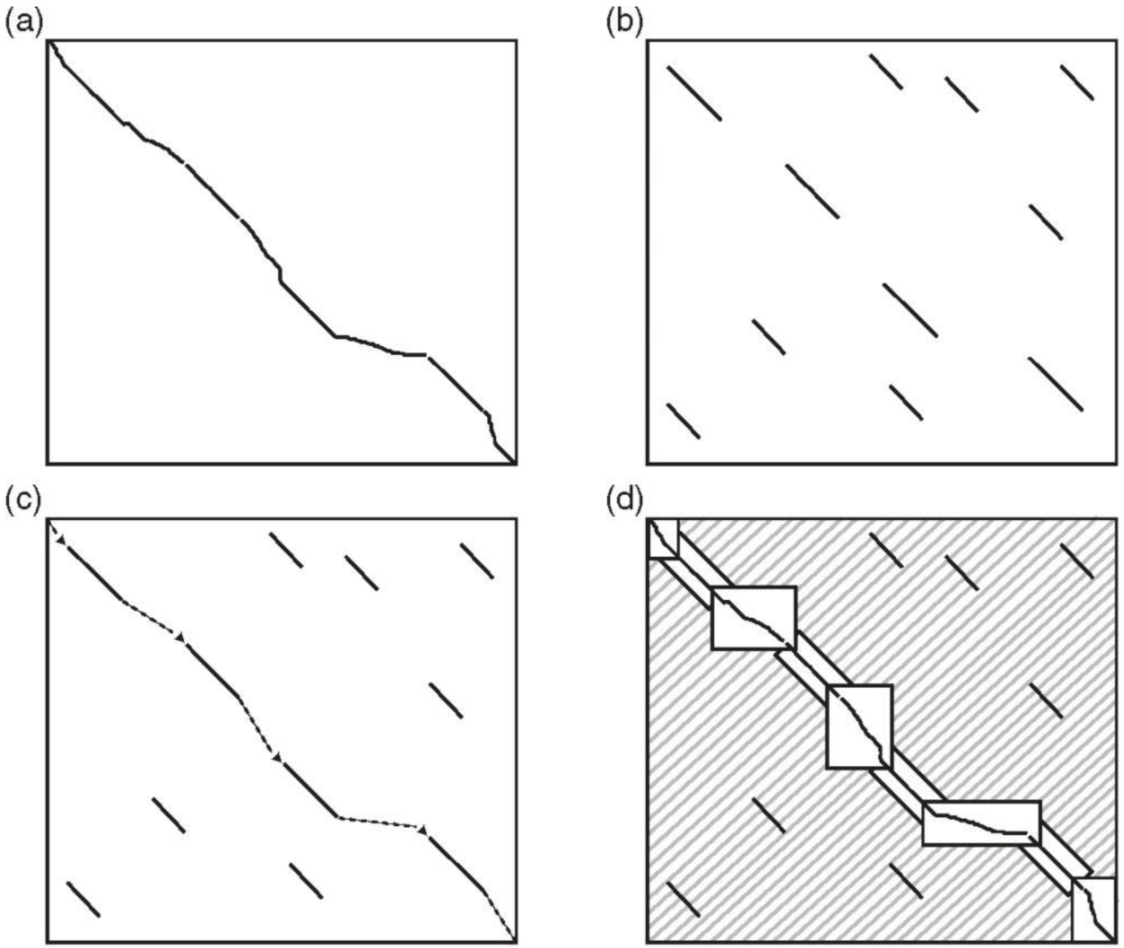
Search tool: BLAT

Description (optional):

Additional configurations: General options ⊞

Run ›

LAGAN (Limited Area Global Alignment of Nucleotides)



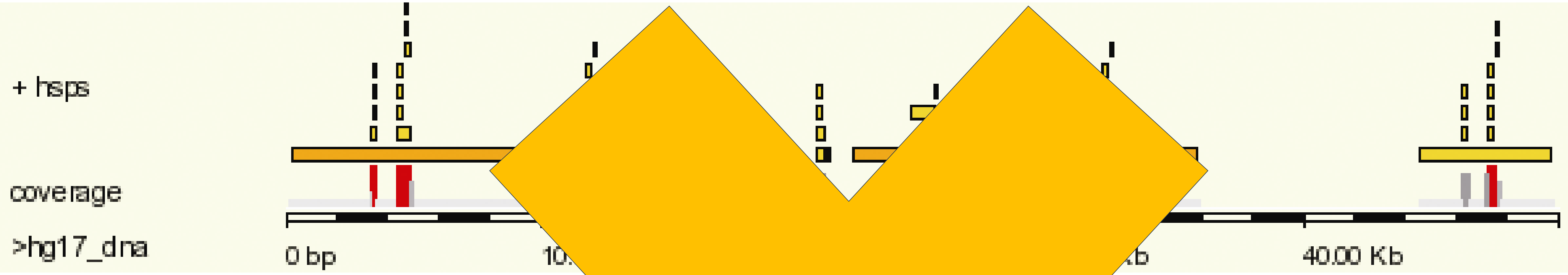
local alignment

identify anchors

join locally aligned segments in chains

compute optimal alignment in boxed areas

SSAHA2



SSAHA converts a DNA da ce sequence such as the human genome) in lected fixed word lengths (*k*-mers). Re d h table for matches by pairwise a