Chapter 5:
Advanced Database Searching

Learning objectives

- 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 PatternHunter, Megablast, BLAT, and LASTZ.

[3] How can we align tens of millions of short reads 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

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

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

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

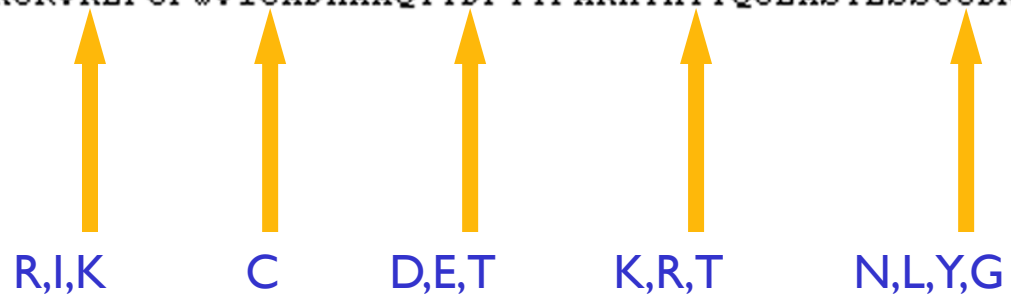
PSI-BLAST is performed in five steps

[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
```



		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	-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

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

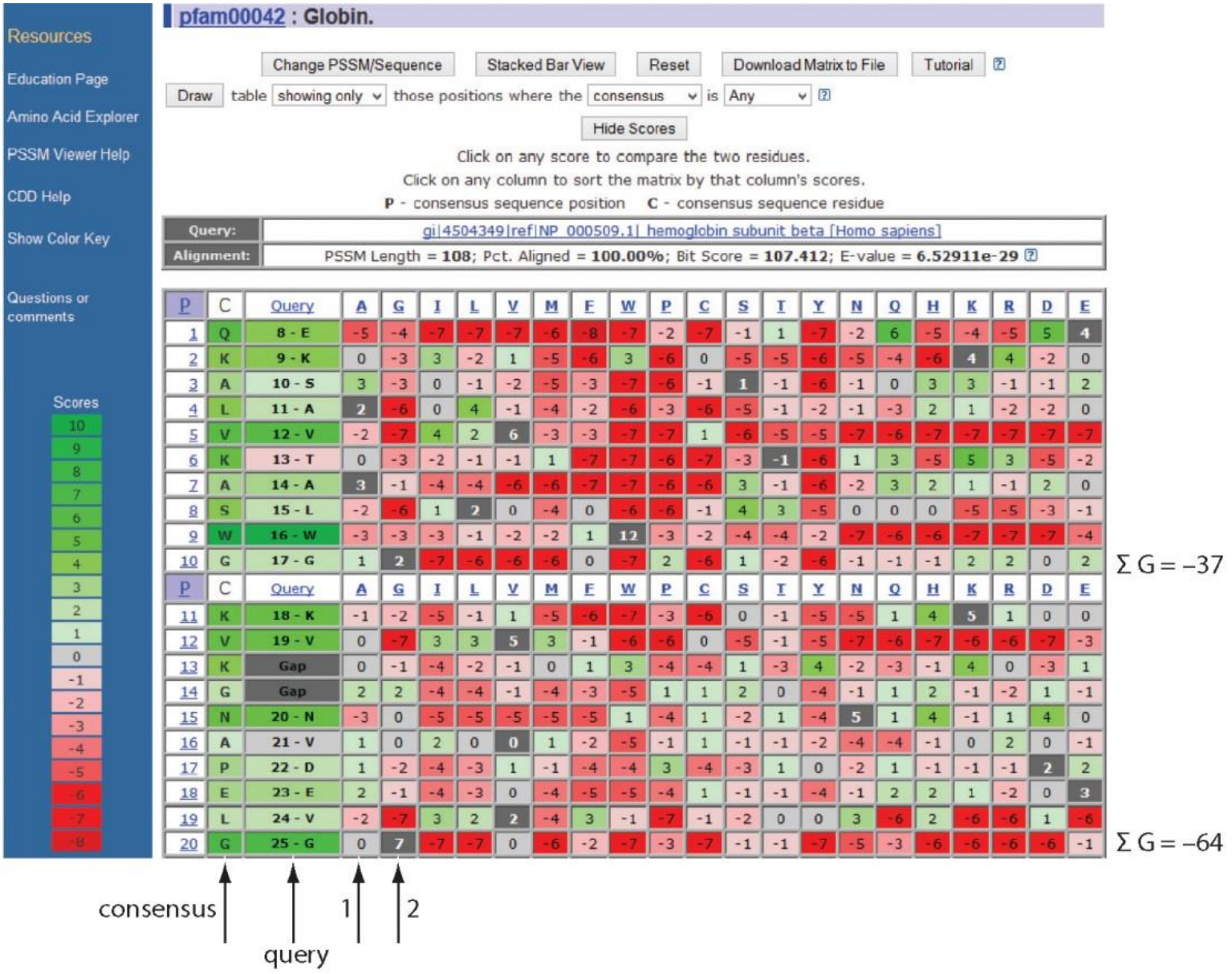
[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)

[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)



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  PSIKHQAANMAGILSLTISQLENLSILDEYLAKLGKLHSRVLNIEEAHFKLMGEAFVQIF  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

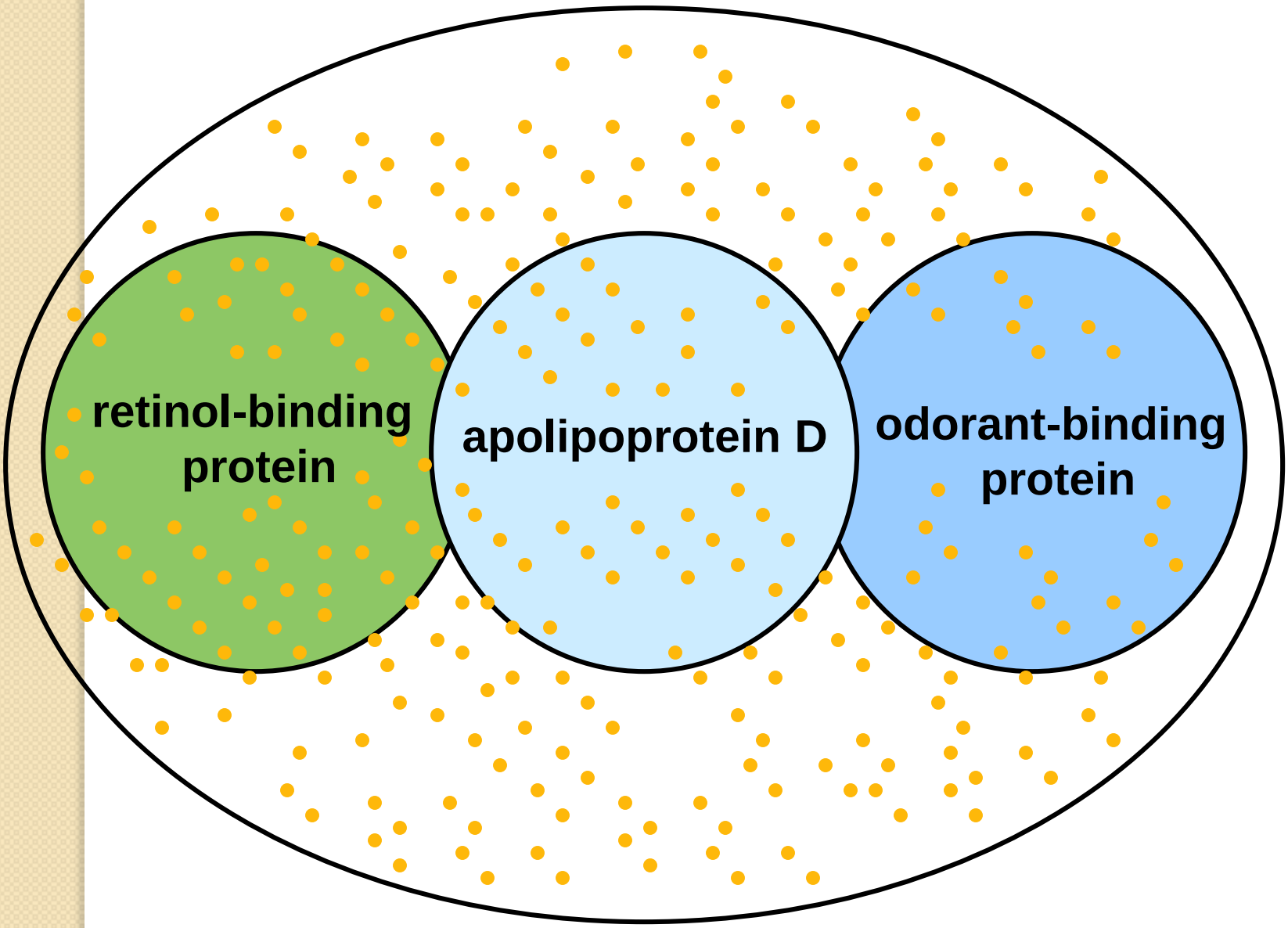
(c) PSI-BLAST iteration 3 (human beta globin versus a *C. albicans* globin)

Range 1: 281 to 426 GenPept Graphics

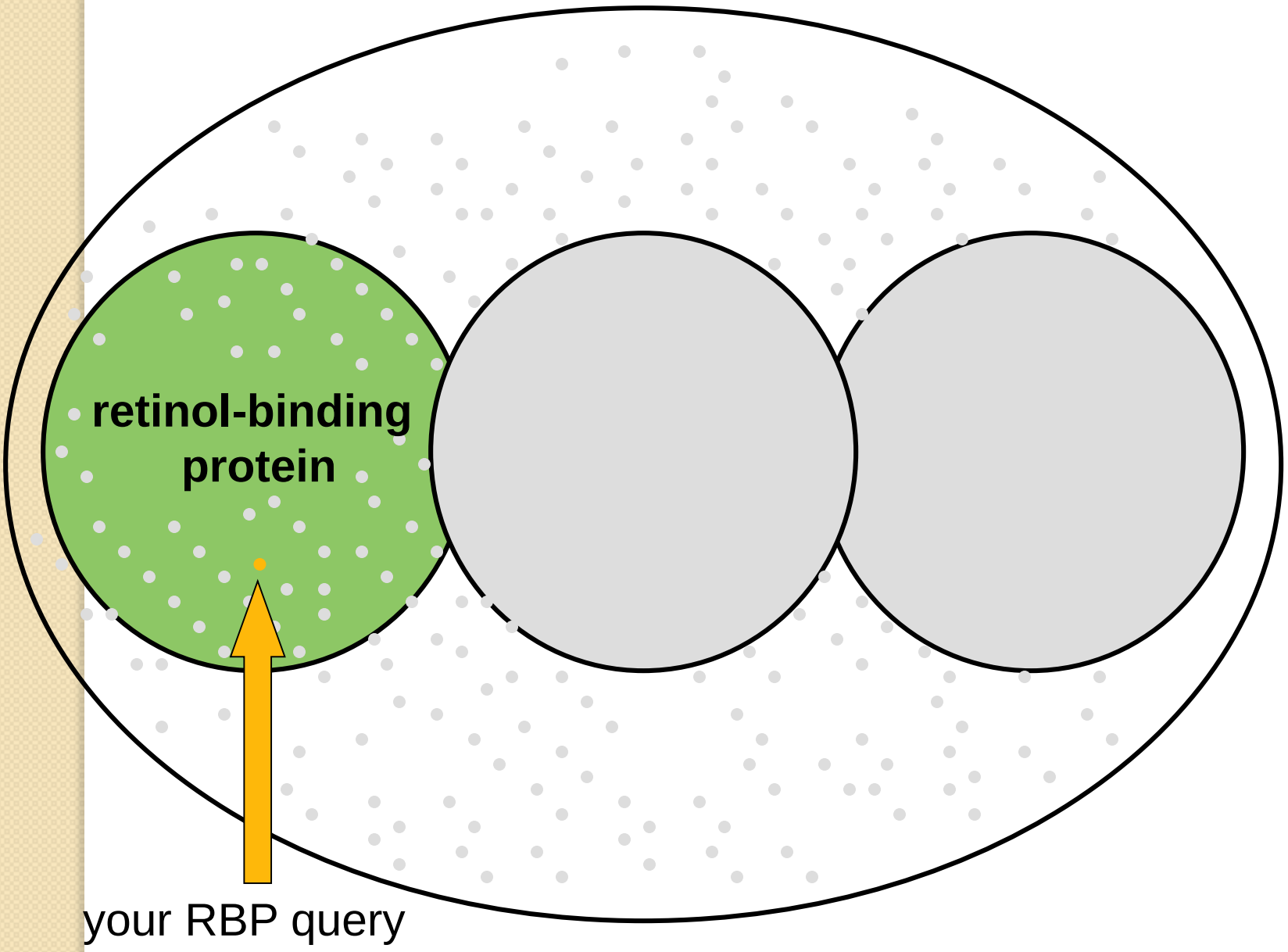
Score	Expect	Method	Identities	Positives	Gaps
128 bits(321)	2e-33	Composition-based stats.	28/146(19%)	50/146(34%)	6/146(4%)

```
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
```

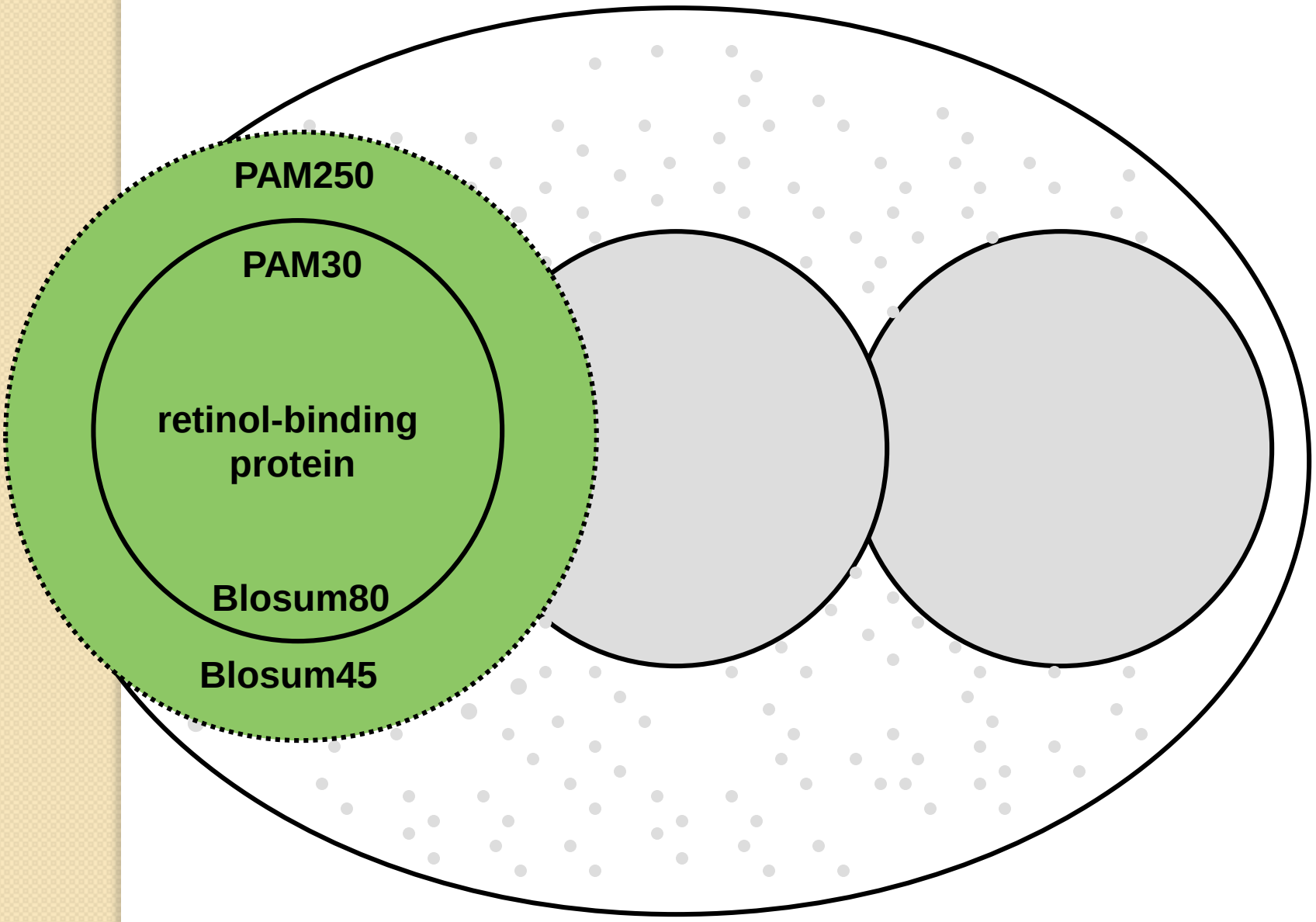
The universe of lipocalins (each dot is a protein)



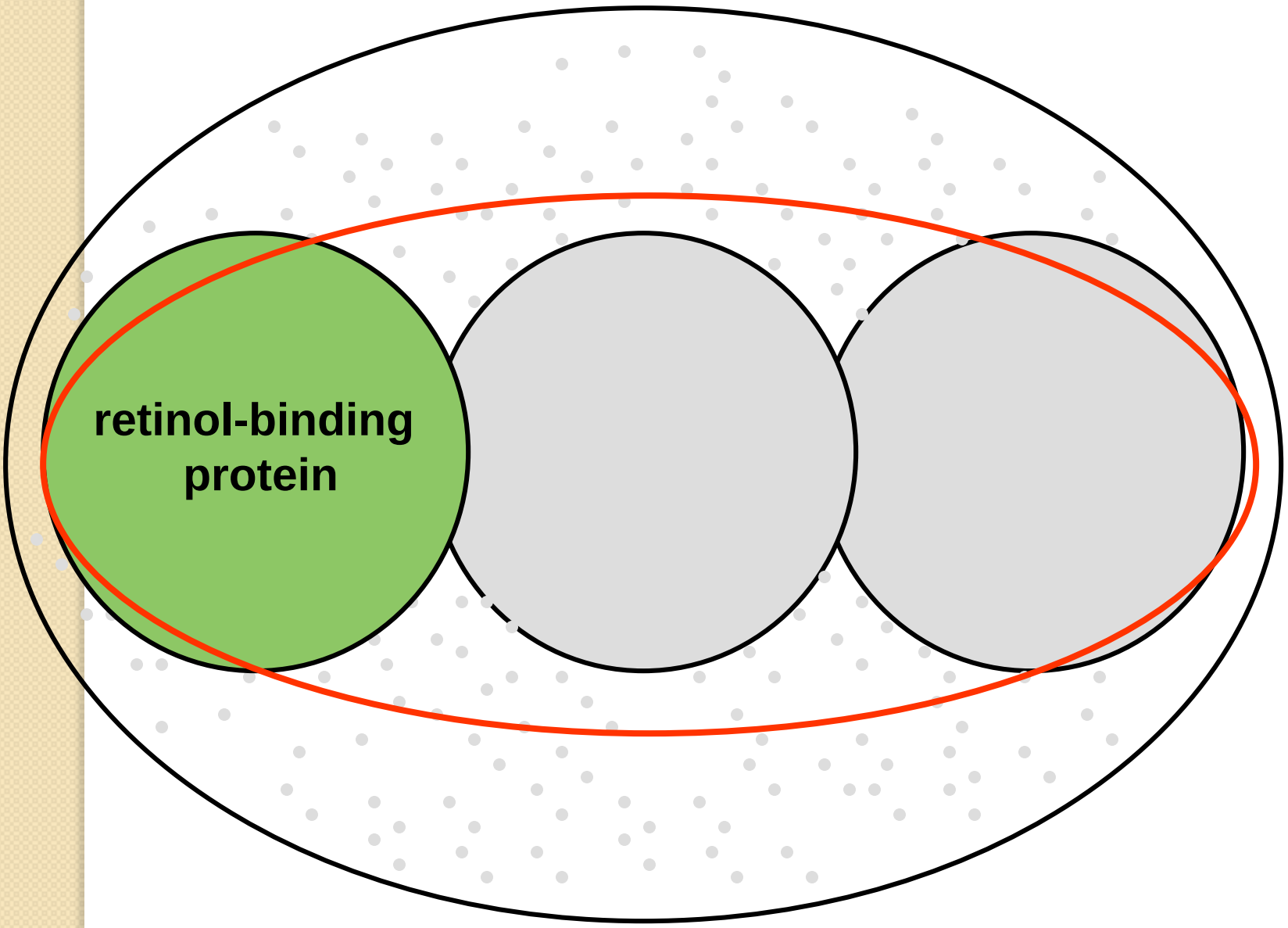
Scoring matrices let you focus on the big (or small) picture



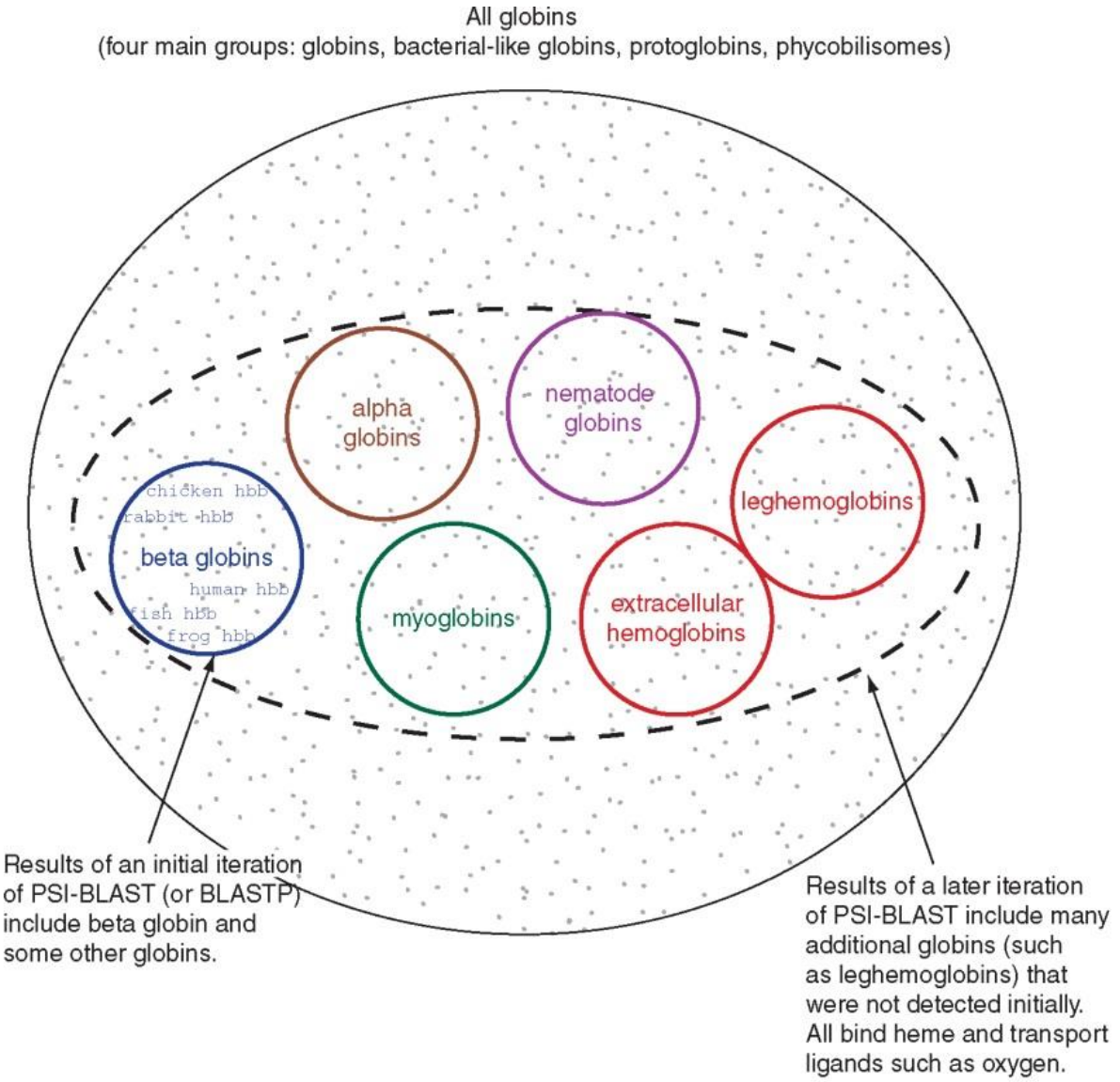
Scoring matrices let you focus on the big (or small) picture



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



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



PSI-BLAST: the problem of corruption

In 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, it may 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

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.

The results are typically superior to those of PSI-BLAST.

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).

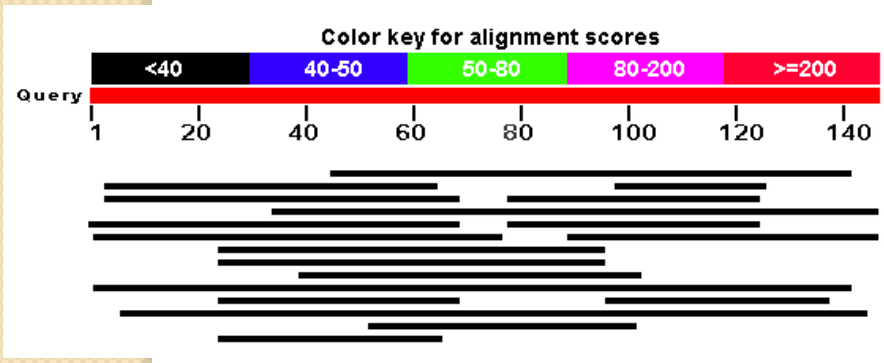
DELTA-BLAST searches a query against a library of pre-computed PSSMs. One reason DELTA-BLAST outperforms PSI-BLAST is that it results in larger, more complete PSSMs than PSI-BLAST.

Most queries do match a PSSM; if not the search proceeds in a PSI-BLAST-like manner.

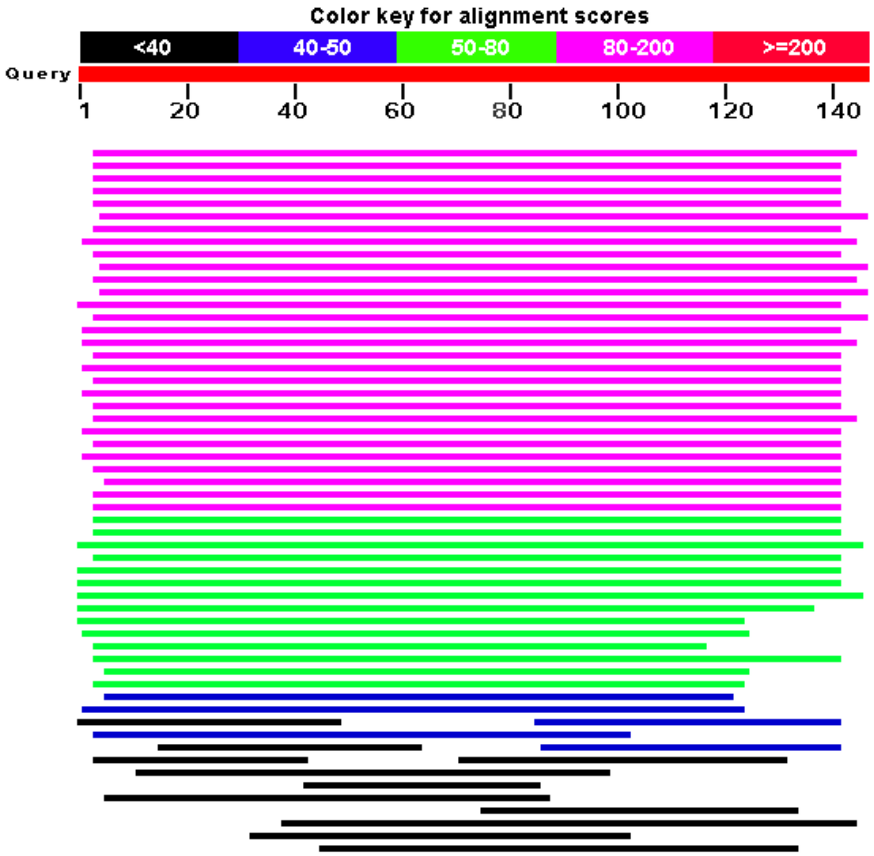
One iteration of DELTA-BLAST is recommended.

Search HBB (NP_000509) against RefSeq plants…

BLAST



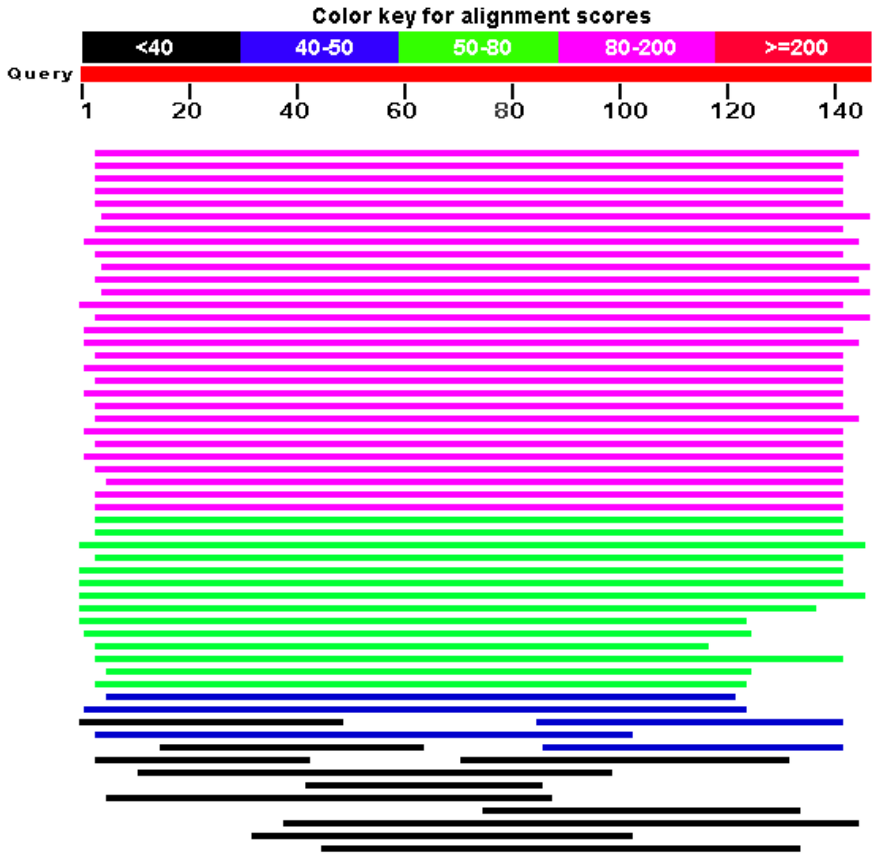
DELTA-BLAST



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



DELTA-BLAST and PSI-BLAST: assessing performance

To assess the performance of BLASTP, PSI-BLAST, DELTA-BLAST or other programs it is necessary to have a "truth" dataset to distinguish true positives, false positives, true negatives, and false negatives.

An approach is to perform searches against databases that incorporate structural information to define homology.

Evaluate PSI-BLAST or other programs' results using a database in which protein structures have been solved and all proteins in a group share ≤ 40% amino acid identity.

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

PHI-BLAST: Pattern hit initiated 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
                        :               *   .   .  .***  .: *.** :*.
```

Inspect an alignment, choose a pattern (manually).

(b) PHI pattern

NFDX(5)GXW[YF]

Follow the rules for the syntax of your pattern.

(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   MVGTFTDIEDPAKFKMKYWG  110
               +  F +  D   FK+ ++G
Sbjct    90   GLAKFVENADTGHFKVSFFG  109
```

The output includes asterisks indicating the position of your pattern.

Try it to boost sensitivity of your search.

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

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.

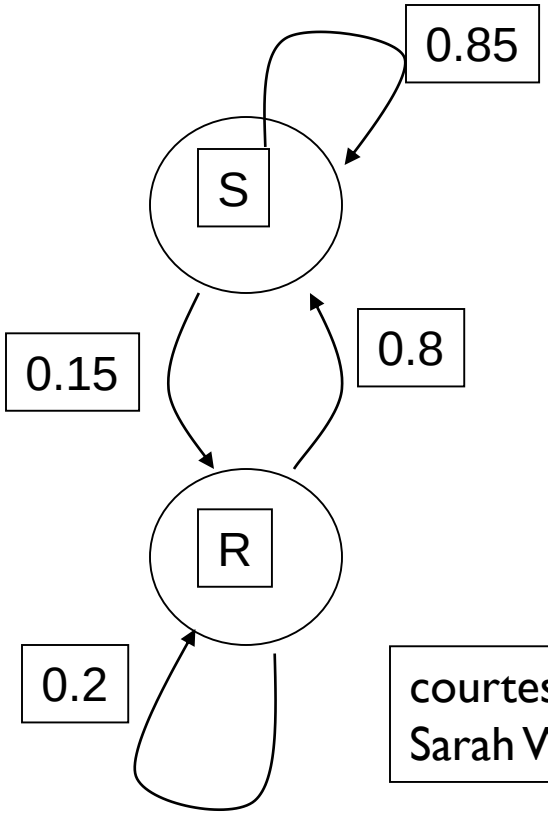
• Hidden Markov models (HMMs) are "states" that describe the probability of having a particular amino acid residue at arranged in a column of a multiple sequence alignment

• HMMs are probabilistic models (unlike DELTA-BLAST and PSI-BLAST)

Simple Markov Model

Rain = dog may not want to go outside

Sun = dog will probably go outside

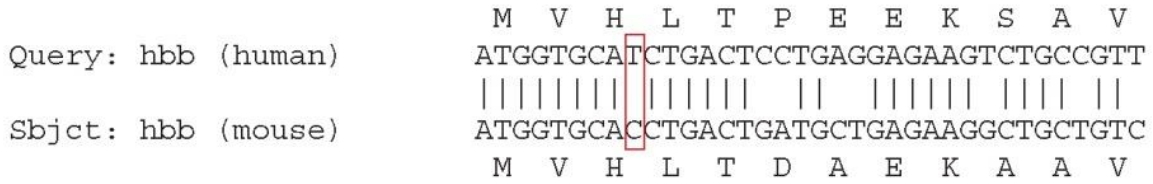


courtesy of Sarah Wheelan

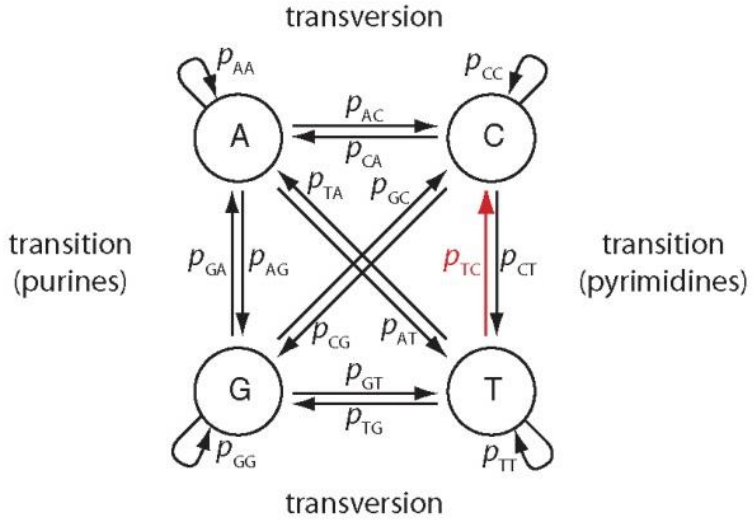
Markov condition = no dependency on anything but nearest previous state ("memoryless")

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

(a)



(b)



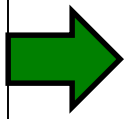
(c)

	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

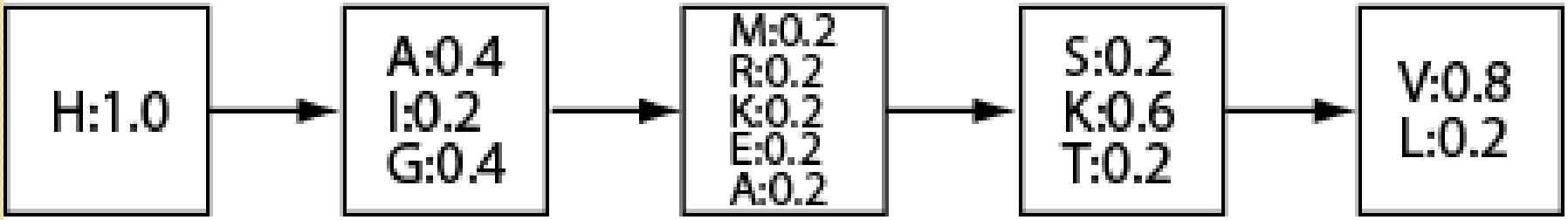


	position				
Probability	1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

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

$$p(\text{HARTV}) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128$$

$$\text{Log odds score} = \ln(1.0) + \ln(0.4) + \ln(0.2) + \ln(0.2) + \ln(0.8) =$$



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)

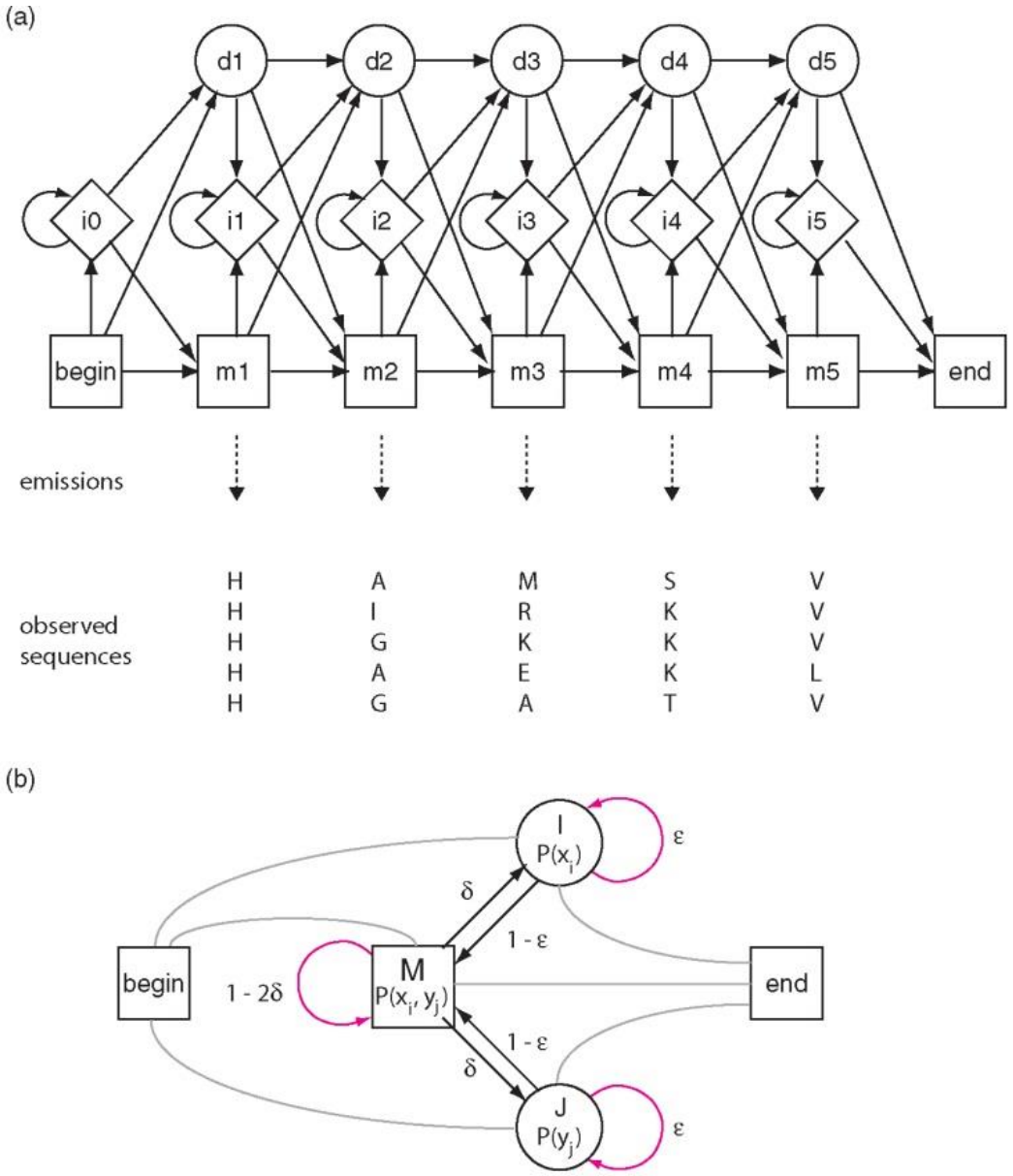
Probability	position 1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

(c)

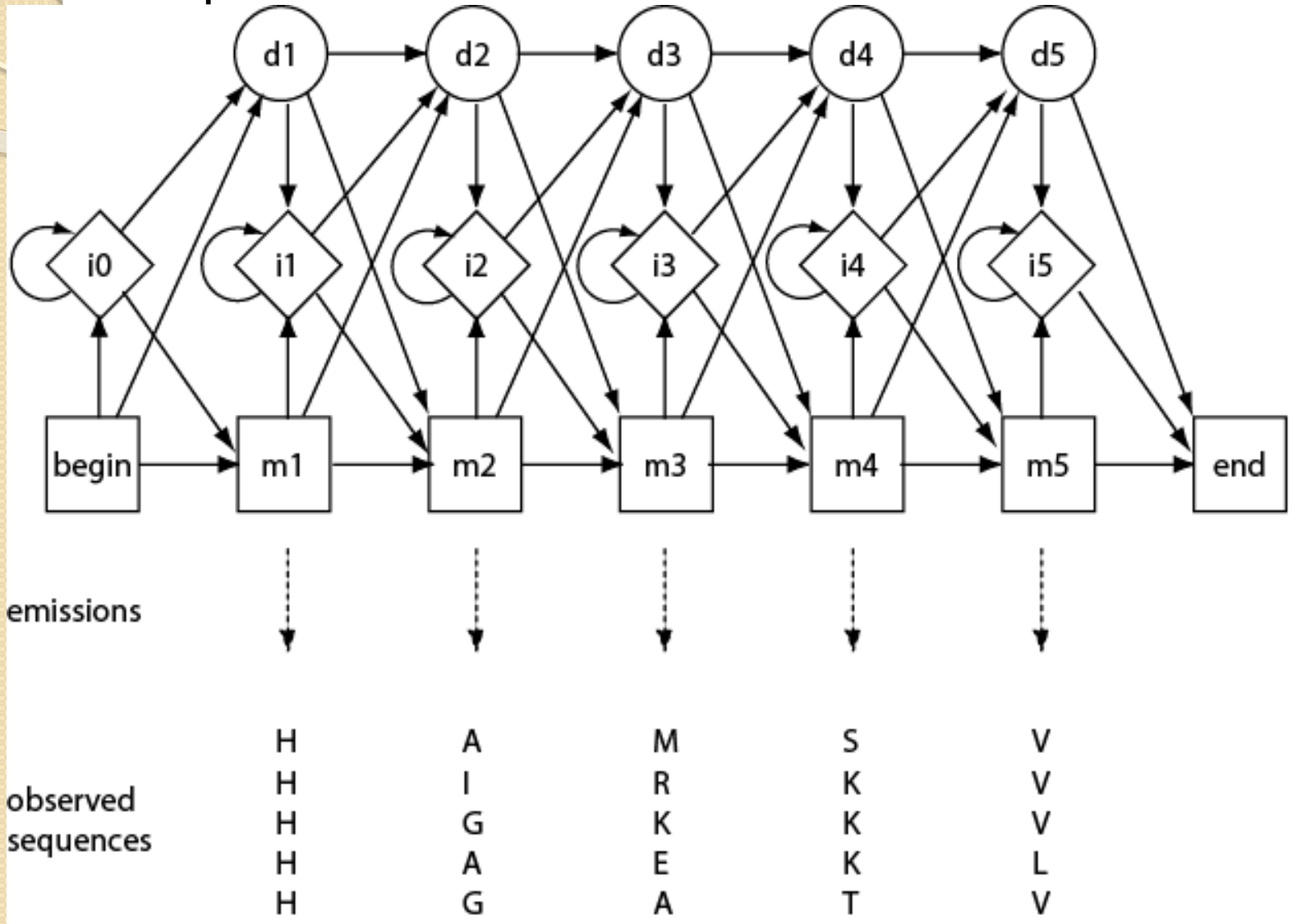
$p(HARTV) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128$

$\text{Log odds score} = \ln(1.0) + \ln(0.4) + \ln(0.2) + \ln(0.2) + \ln(0.8) = -4.357$

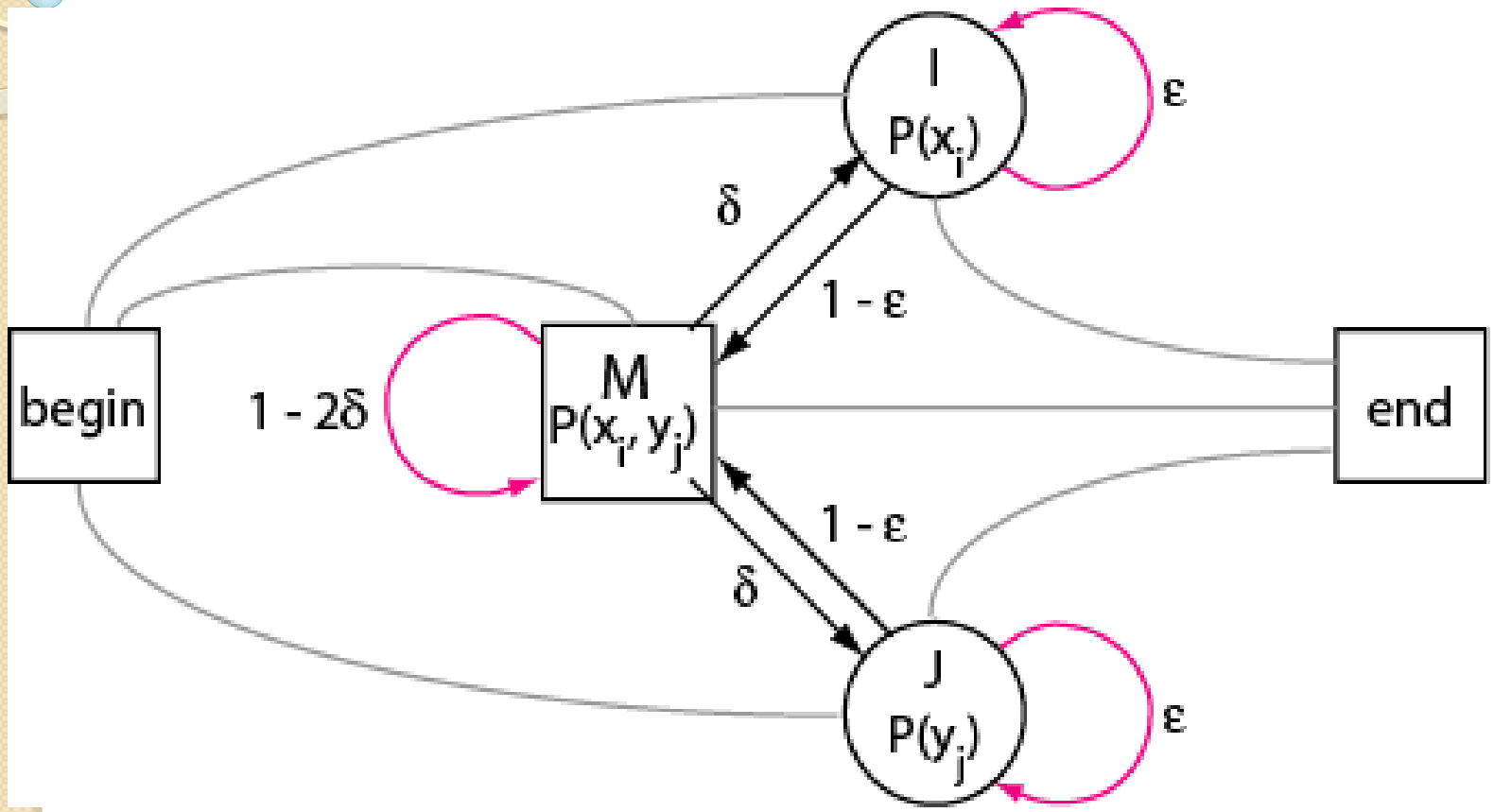
(d)


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

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



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

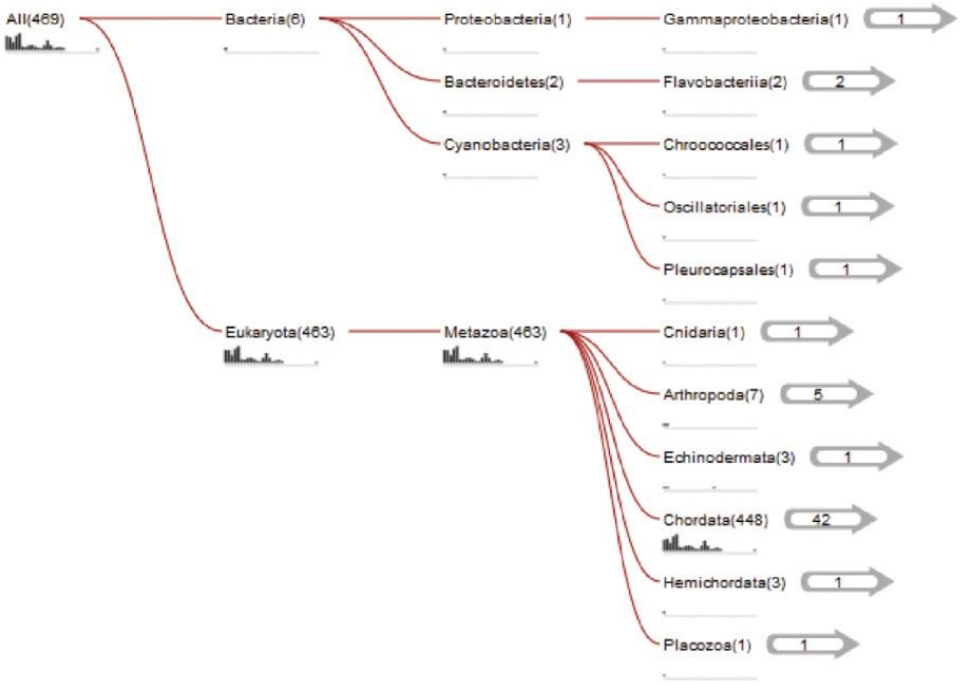
```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# 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

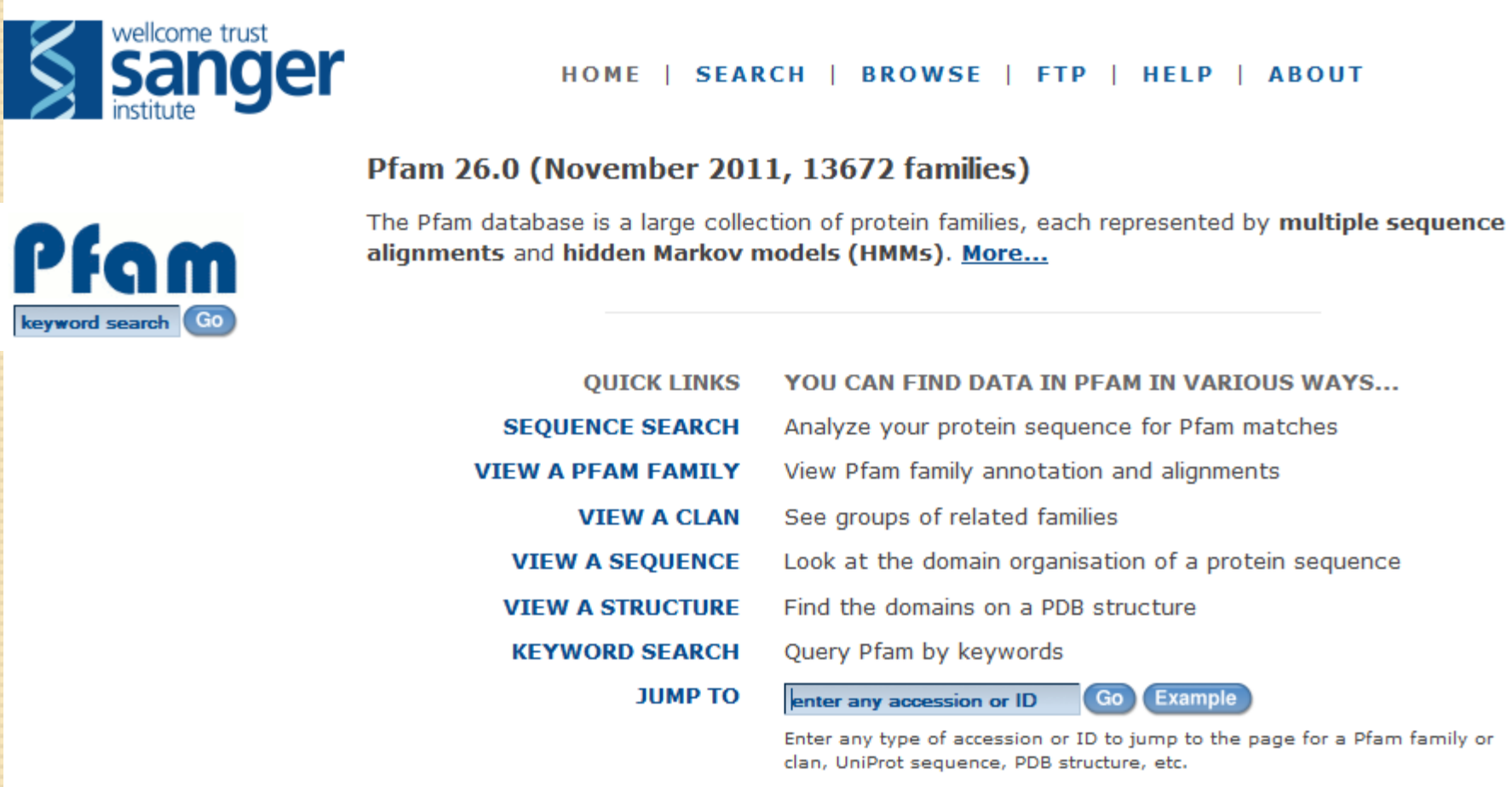
(a) HMMER web output

(b) HMMER phylogenetic output

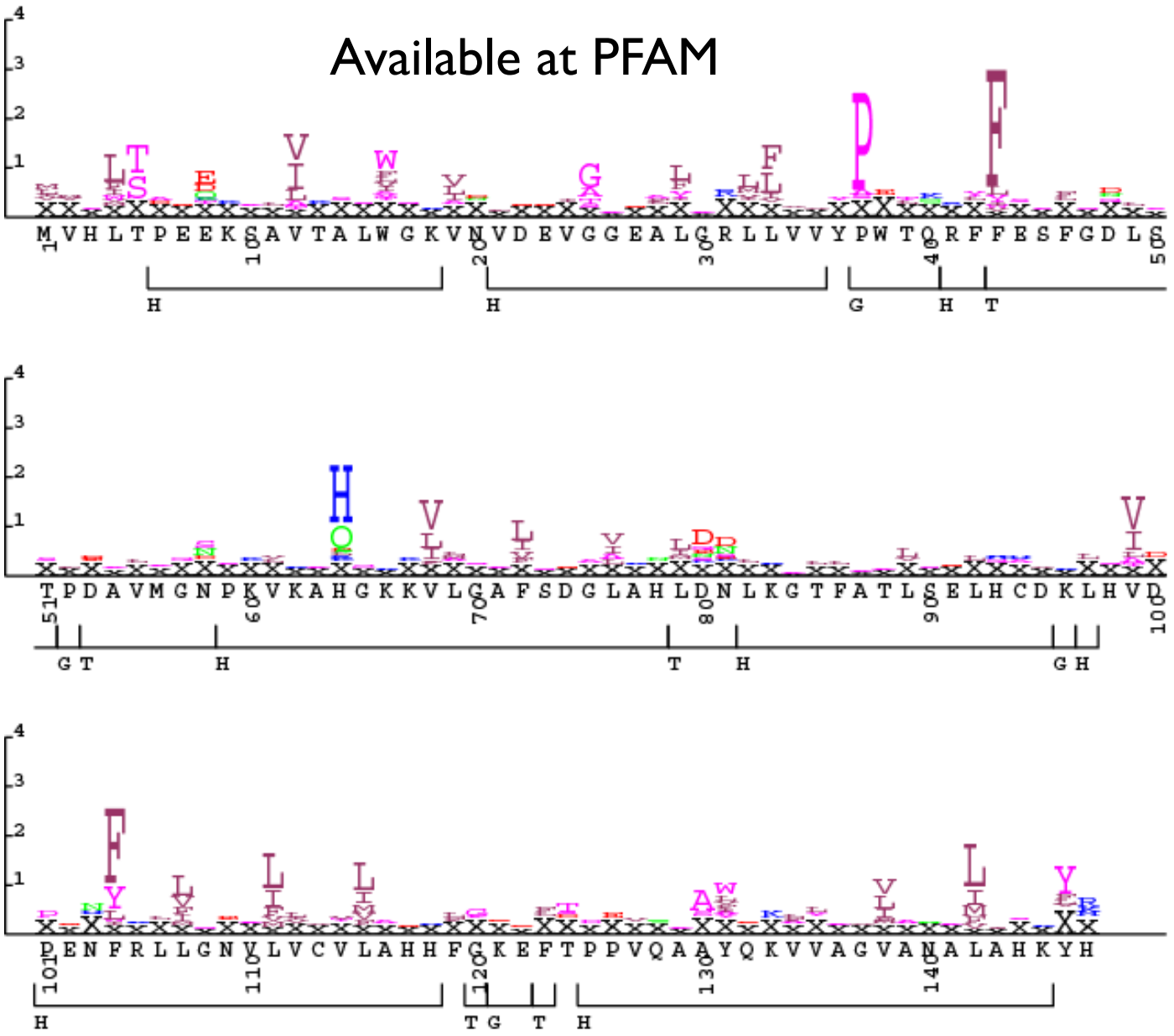


PFAM is a database of HMMs and an essential resource for protein families

http://pfam.sanger.ac.uk/



HMM logos graphically depict the likelihood of observed amino acids



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

BLAST-related tools for genomic DNA

The analysis of genomic DNA presents special challenges:

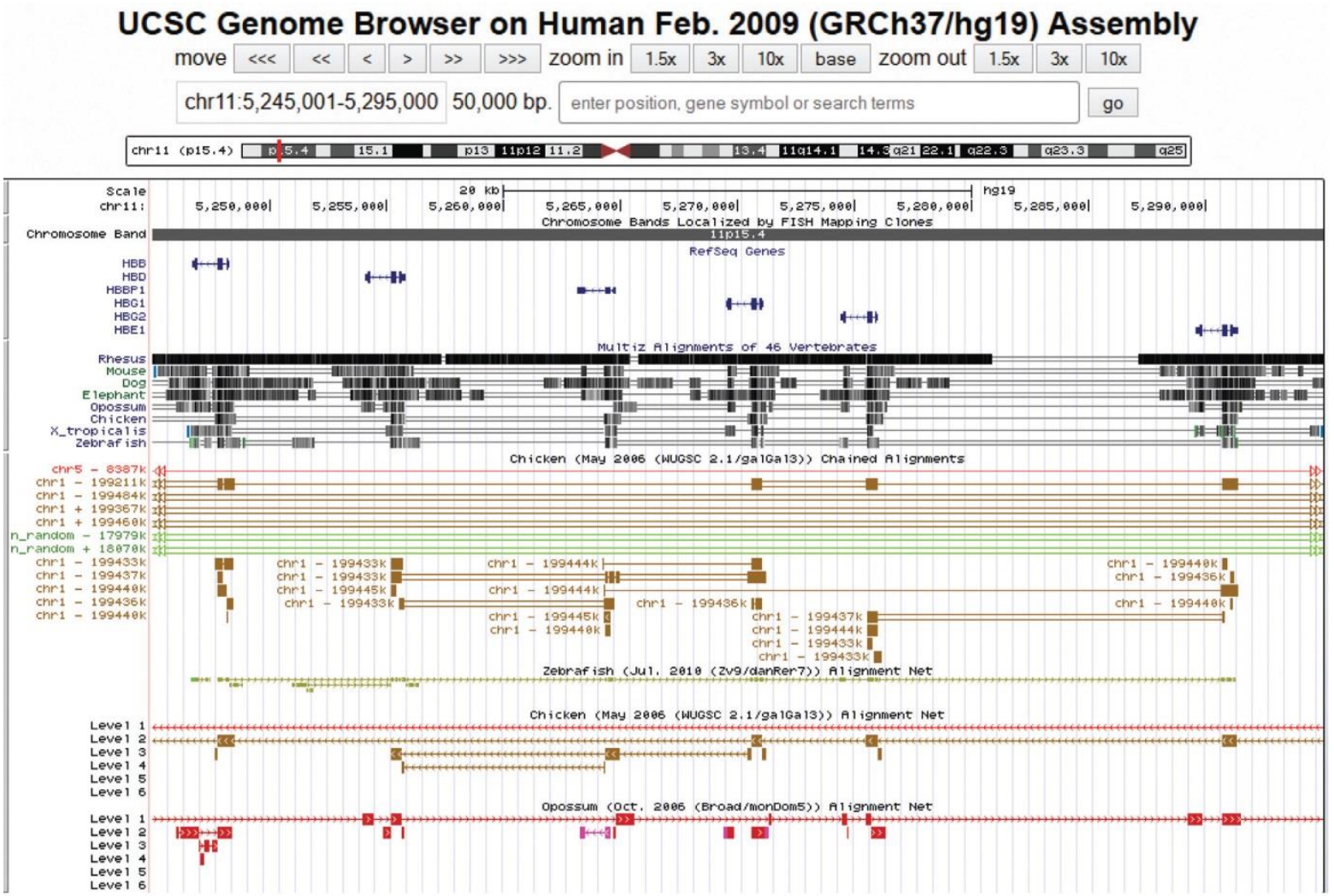
- There are exons (protein-coding sequence) and introns (intervening sequences).
- There may be sequencing errors or polymorphisms
- The comparison may between be related species (e.g. human and mouse)

BLAST-related tools for genomic DNA

Recently developed tools include:

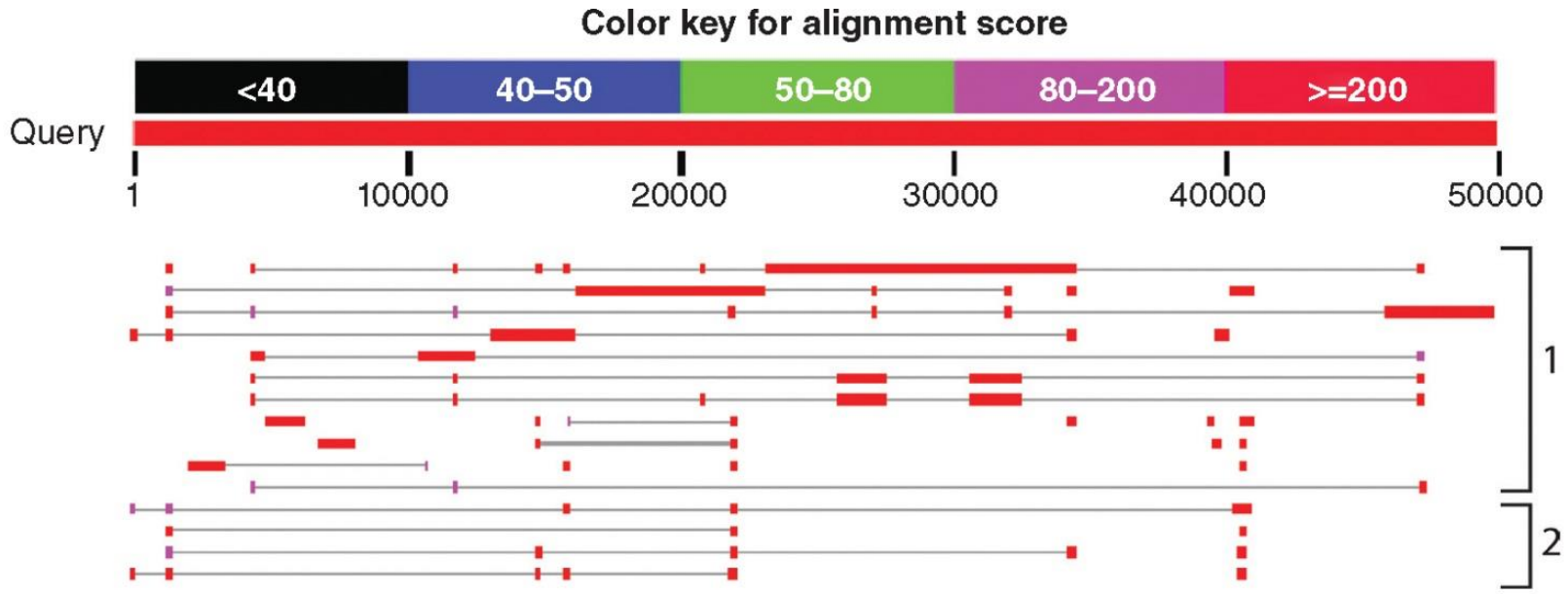
- MegaBLAST at NCBI.
- BLAT (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into words (11mers), then searches them against a query. Thus it is a mirror image of the BLAST strategy. See http://genome.ucsc.edu
- SSAHA at Ensembl uses a similar strategy as BLAT. See http://www.ensembl.org

BLASTZ alignments at UCSC



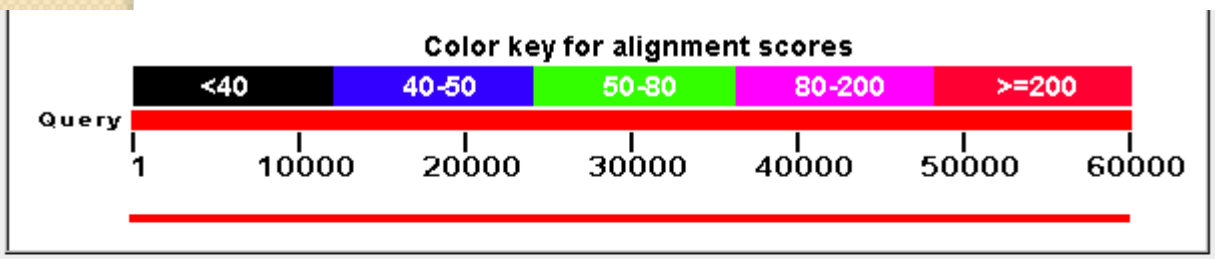
50 kilobases at the beta globin locus are displayed, including BLASTZ alignments.

MegaBLAST: extremely fast searches with large seeds

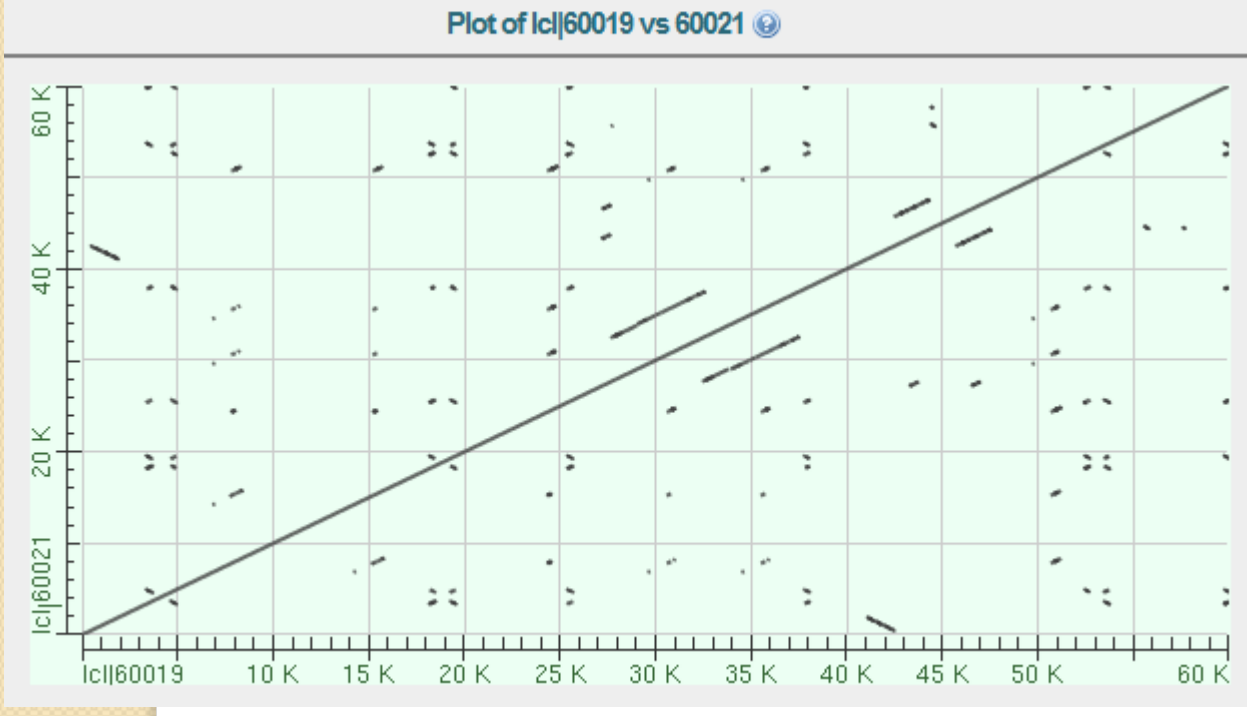


--very fast
--uses very large word sizes (e.g. w=28, up to w=256)
--use it to align long, closely related sequences
--Choose discontiguous megablast for cross-species comparisons (tolerates mismatches)

MegaBLAST output

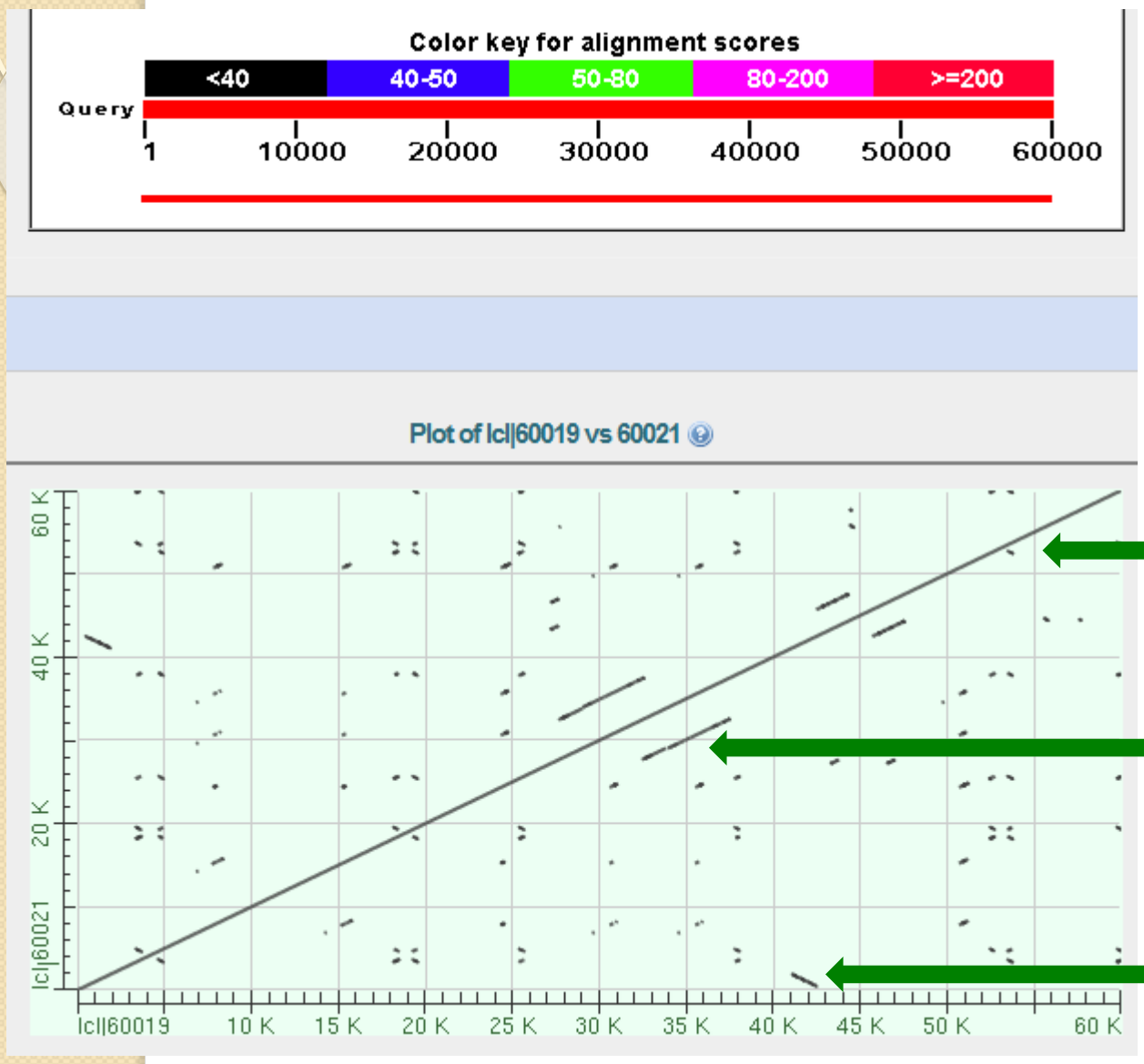


60 kb of the human beta globin locus versus itself!



Dot matrix view

MegaBLAST output



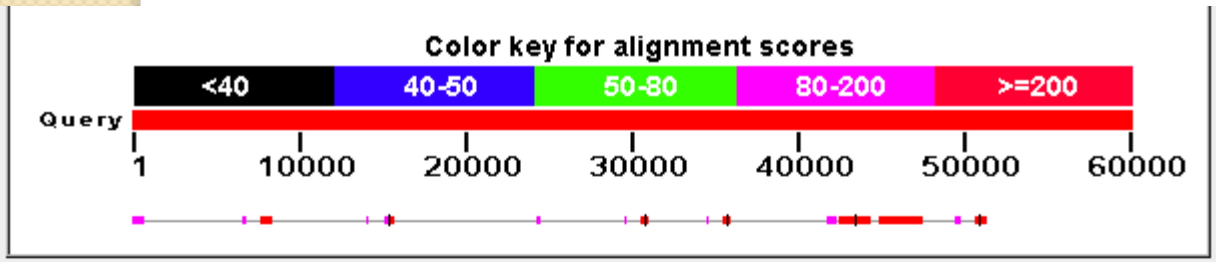
60 kb of the human beta globin locus versus itself!

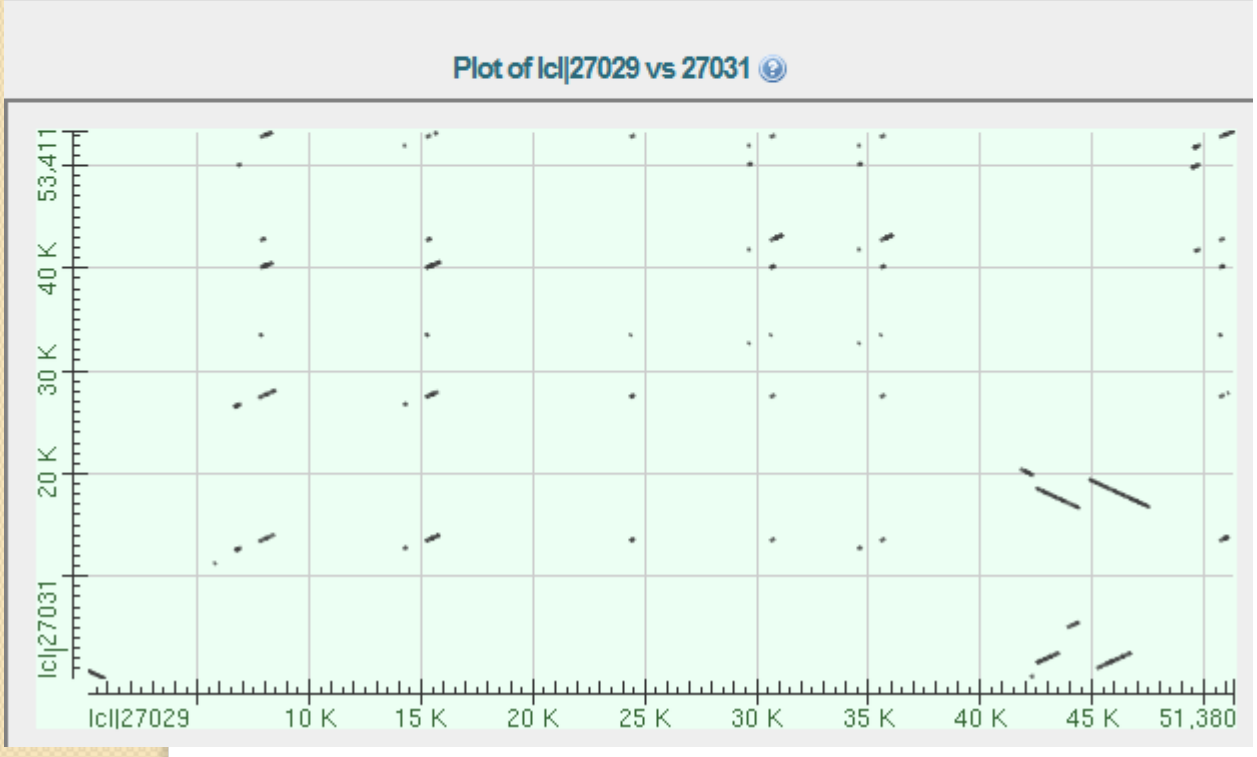
main diagonal reflects exact matches

off diagonal reflects duplications

Flipped, off diagonal reflects inversions

MegaBLAST output





60 kb of the **human** beta globin locus versus 60 kb of the **mouse** beta globin locus

BLAT indexes a whole genomic database rather than a query

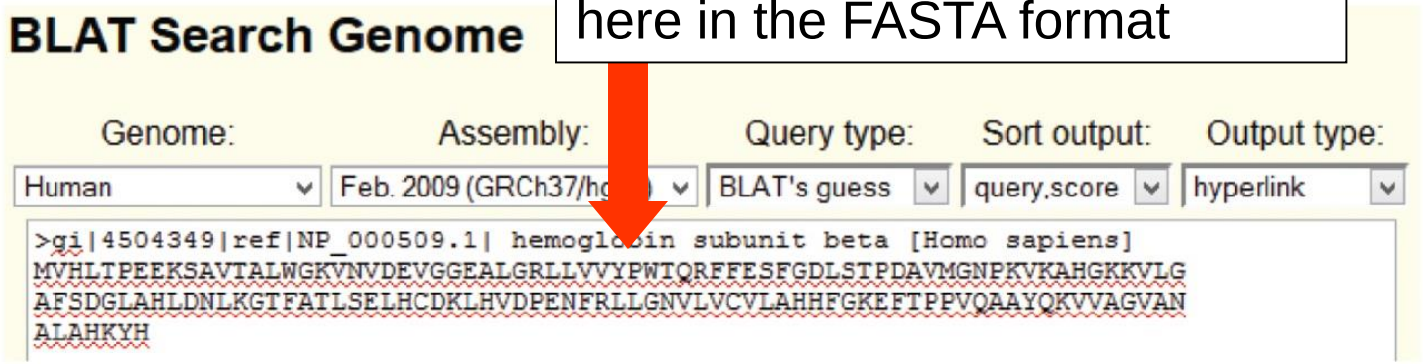
"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80%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

--BLAT website

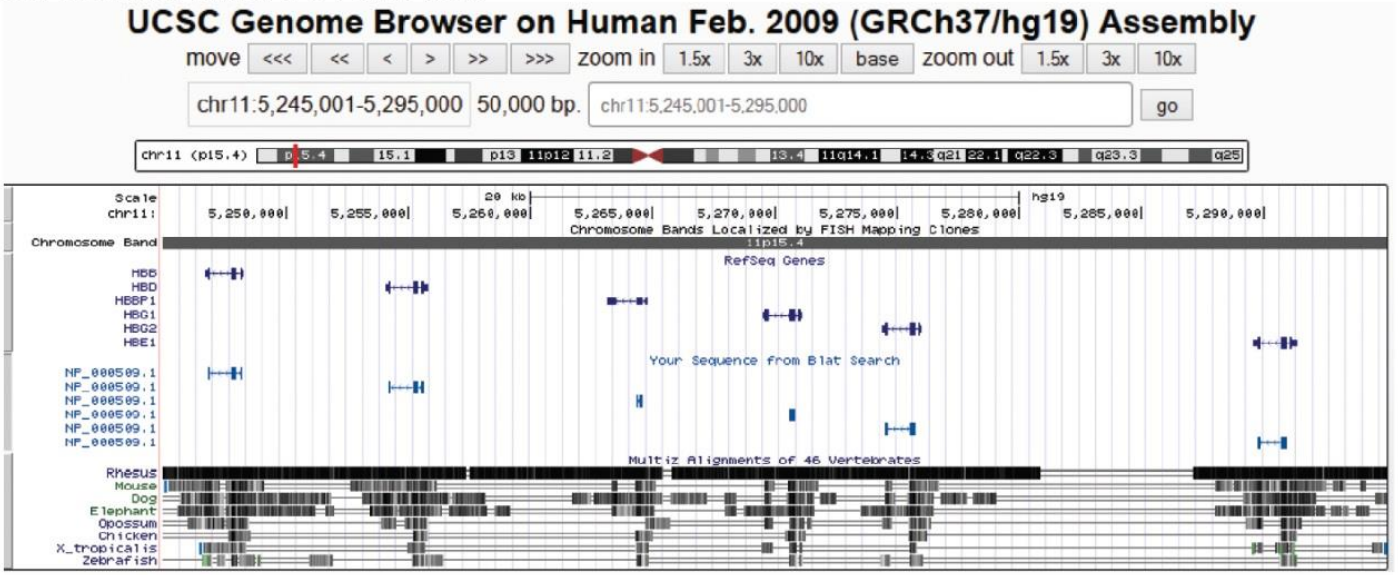
BLAT indexes a whole genomic database rather than a query

(a) BLAT query (protein or DNA)

Paste DNA or protein sequence here in the FASTA format



(b) BLAT result (zoomed to 50 kilobases)



BLAT output includes browser and other formats

Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help

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

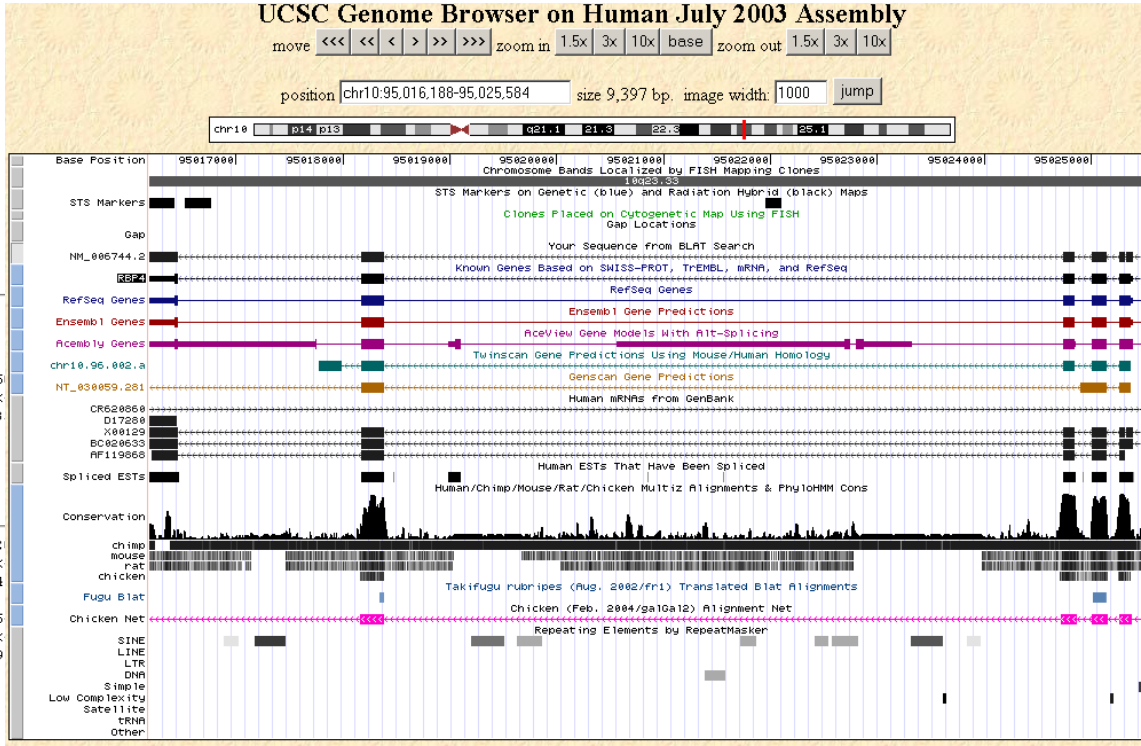
Alignment of NM_006744.2

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

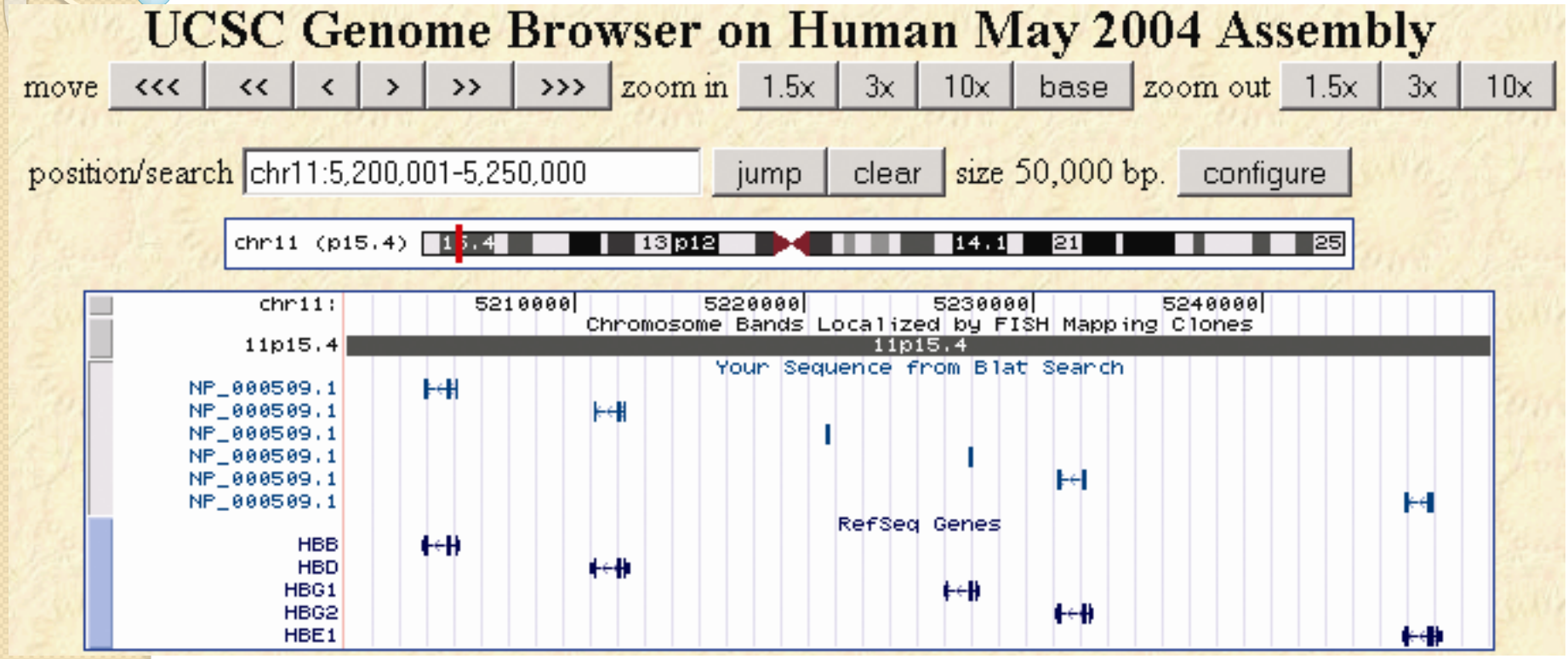
```
ttggtcacga actgacccac tacaccaaac agatgggaac ttagaccaaa  95017085
atcaaattga ttttgaagct tcacccttca aaattaaatg tagggccagc  95017035
aaggattcca ccacagcagg gcaccgtagg aagcagatta tgcacattgt  95016985
acagatgaga aaacagaggc tgacaatgag gaagaaagtt gccctaaatc  95016935
tttaataggt agctgagttg gatgcaaacc tagagttttg 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 0000005
>>>>>>>> |||||||||||||||||||||||||||||||||| |||||||||||||||| <<<<<<<
95025584 cgctcgcctccctcgctccacgcgcgcccggactcggcggccaggcttgc 9502553

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

00000071 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 0000012
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
95025390 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 9502534

00000121 ggcg.tgggcagcggcc.....gagcgcgactgccgagtgagcagcttcc 0000016
>>>>>>>> |||| |||||||||||||     |||||||||||||||||||||||||| <<<<<<<
95025340 ggcgctgggcagcggccgcgcggagcgcgactgccgagtgagcagcttcc 9502529

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



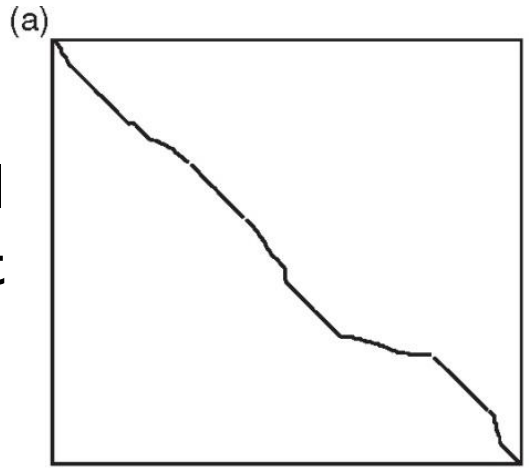
BLAT output includes browser and other formats



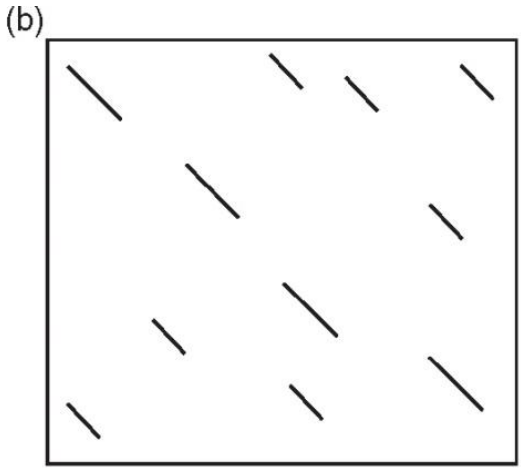
This example shows a BLAT query of beta globin resulting in a series of matches to homologous, neighboring globins.

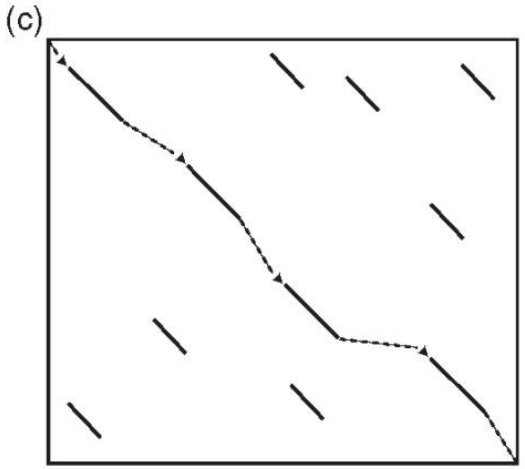
LAGAN (Limited Area Global Alignment of Nucleotides)

local alignment

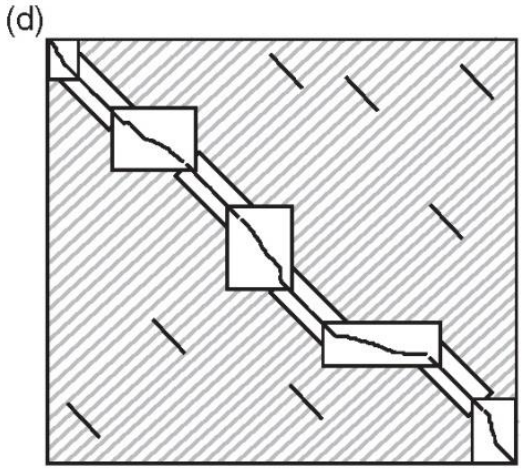


identify anchors



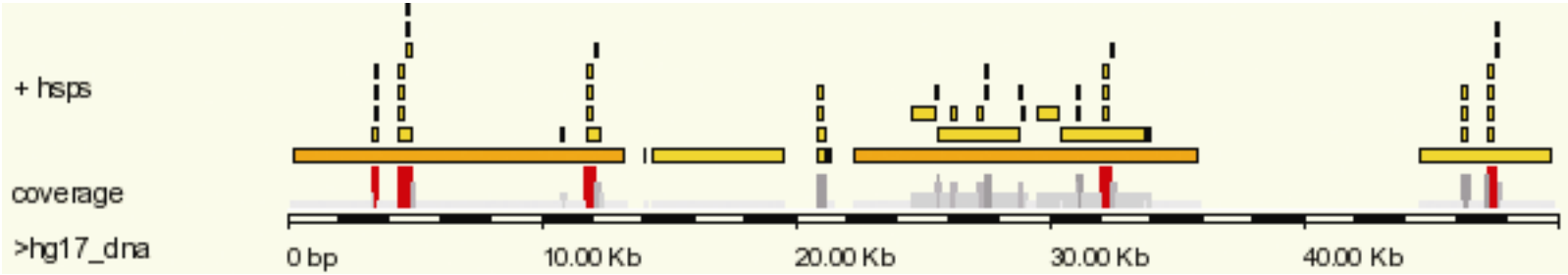


join locally aligned segments in chains



compute optimal alignment in boxed areas

SSAHA



SSAHA converts a DNA database (with a reference sequence such as the human genome) into a hash table with user-selected fixed word lengths (*k*-mers). Reads are searched against this hash table for matches by pairwise alignment.

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

Next-generation sequencing (NGS)

In NGS, many millions of short reads (~150 base pairs) must be aligned to a long reference (~3 billion base pairs).

BLAST would be unacceptably slow.

Current aligners use BLAST-related strategies such as hash tables, gapped and ungapped alignment, and long seeds.

Sequence alignment for next-generation sequencing

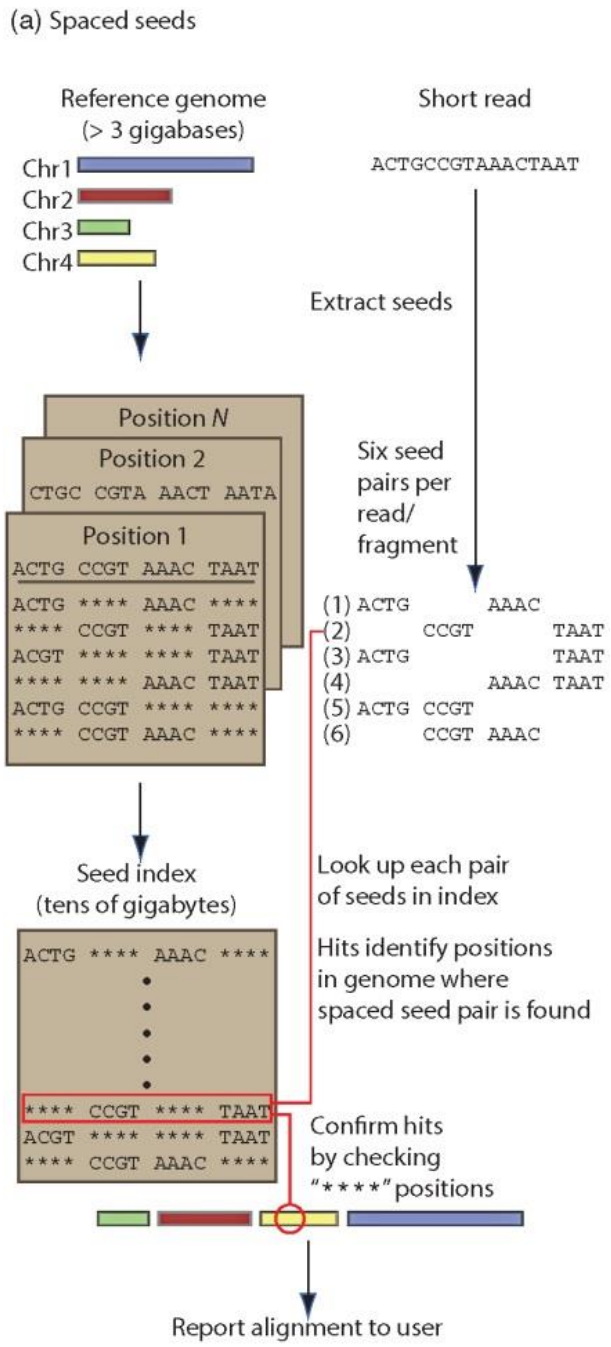
Applications include:

- DNA sequencing (e.g. re-sequencing a human genome)
- RNAseq (measuring RNA transcript levels)
- ChIP-seq (finding protein binding sites)
- Methylation studies (genome-wide)

Two approaches to sequence alignment for NGS

- [1] **Hash Tables**
  --BLAST seed and extension approach uses hash table indexing
  --spaced seed aligners index the reads or the genome
  --programs vary the number of spaced seeds, the read length, memory usage, and sensitivity requirements
  --some require multiple seed matches
  --some allow gaps

  [2] **Suffix trees**
  --Step 1: identify exact matches
  --Step 2: build alignments supported by exact matches
  --Example: MUMmer (maximal unique matches)
  --Example: Bowtie and BWA

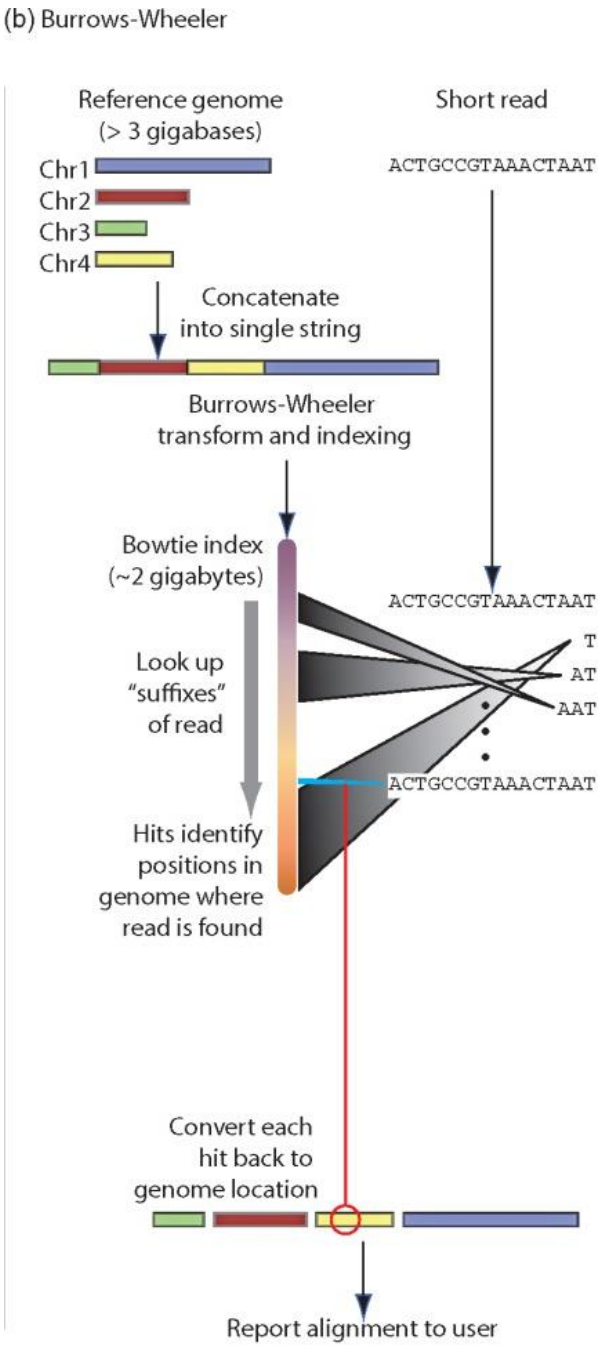


Spaced seed strategy for alignment of many short reads to a large reference genome

Maq uses spaced seed indexing

- Cut the reference genome into "seeds"
- Store seeds in look-up table
- Cut each read into seeds
- Allow up to 2 mismatches in seed pairing reads to the reference

PMID: 19430453

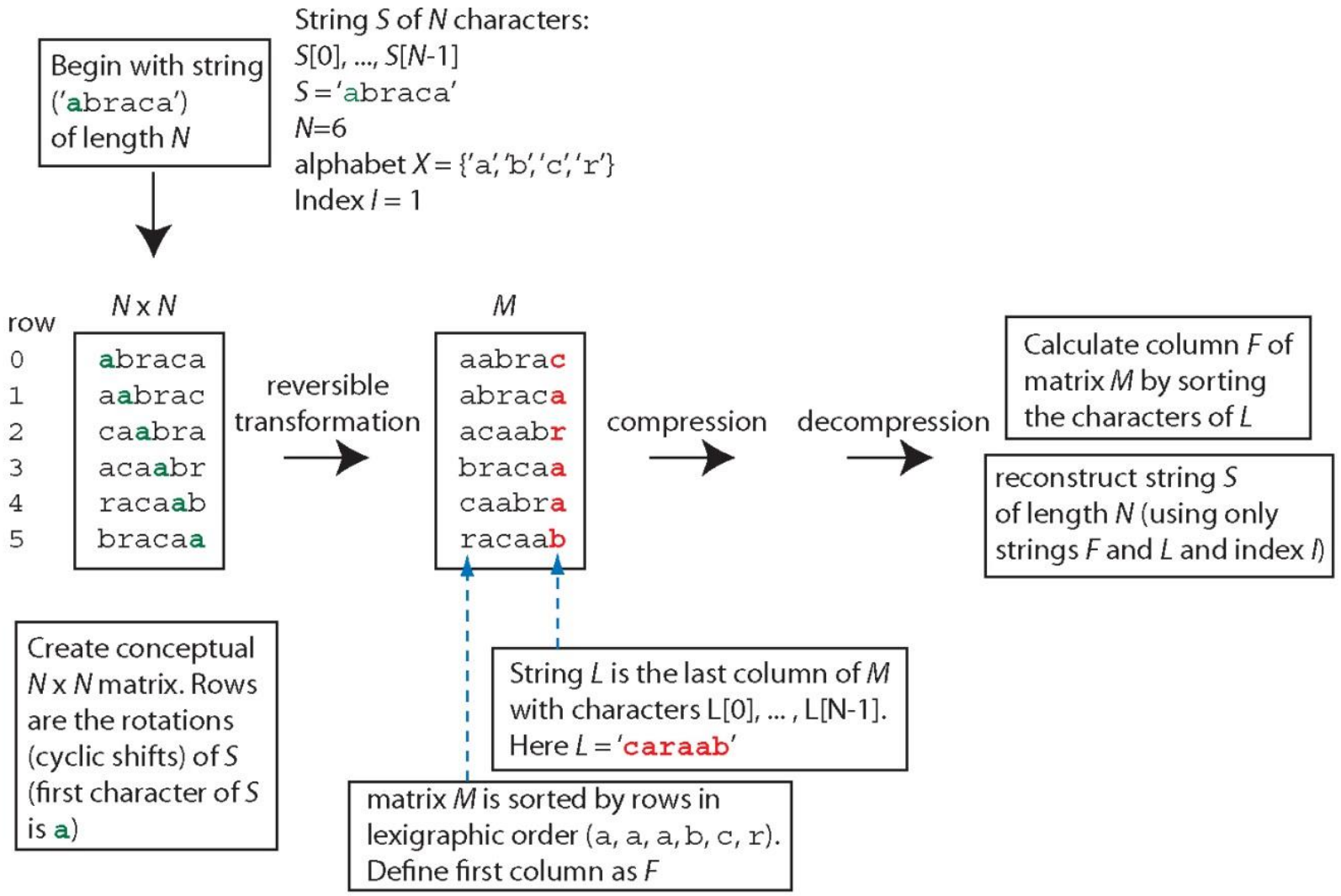


Burrows-Wheeler strategy for alignment of many short reads to a large reference genome

Bowtie uses the Burrows-Wheeler transform (BWT)

- Create a memory-efficient representation of the reference genome (need <2 GB memory; Maq may require >50 GB)
- Align a read one base at a time to a BWT transformed genome
- Progressively solve the alignment of 1 character, then 2, then 3, etc.
- 30-fold faster than Maq

Burrows-Wheeler Transform (BWT): a string (e.g. a reference genome) is compressed then decompressed to facilitate alignment



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

Perspective

- For database searching there continue to be many innovative approaches to improve sensitivity and specificity.
- We discussed DELTA-BLAST which is usually the best algorithm for any protein search.
- For DNA searches innovative spaced seed approaches have greatly increased search speed.
- For next-generation sequence data, many algorithms have been introduced to align a vast number of reads (e.g. 1 billion short reads) to a large reference genome (e.g. the ~3 billion base pair human reference genome).