Supergenomic Network Compression and the Discovery of EXP1 as a Glutathione Transferase

Andreas Martin Lisewski,[1,2,13,*] Nagireddy Putluri,[4,5] Richard T. Manuel Llinás,[11,12] Arun Sreeku
[1]Department of Molecular and Huma
[2]Computational and Integrative Biom
[3]Integrative Molecular and Biomedica
[4]Department of Molecular and Cell B
[5]Verna and Marrs McLean Departme
[6]Department of Biochemistry and Mc
[7]Department of Pharmacology
Baylor College of Medicine, Houston
[8]Department of Microbiology and Imi
[9]Division of Infectious Diseases, Dep
Columbia University Medical Center,
[10]Laboratory of Malaria and Vector R
MD 20852, USA
[11]Department of Molecular Biology a
[12]Department of Biochemistry and M
State College, PA 16802, USA
[13]Co-first author
*Correspondence: lisewski@bcm.edu
http://dx.doi.org/10.1016/j.cell.2014.

nding histogram of sequence identities to the query sequence and a dendrogram of true matches across species for

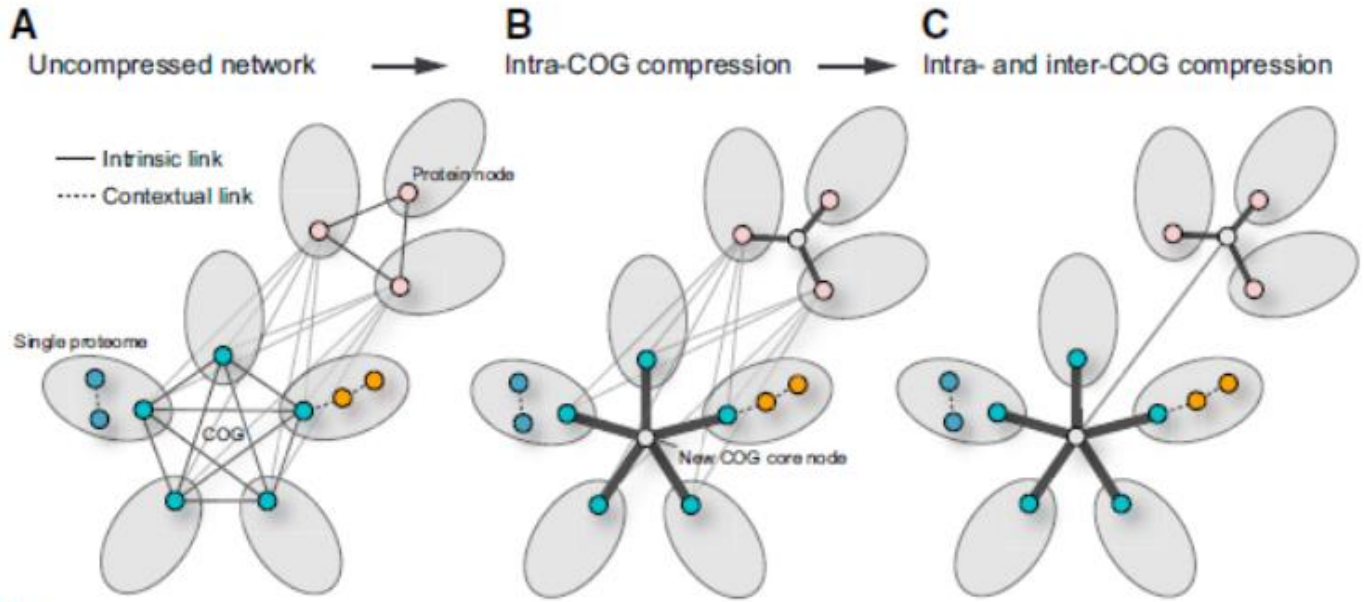
ores cut-off.

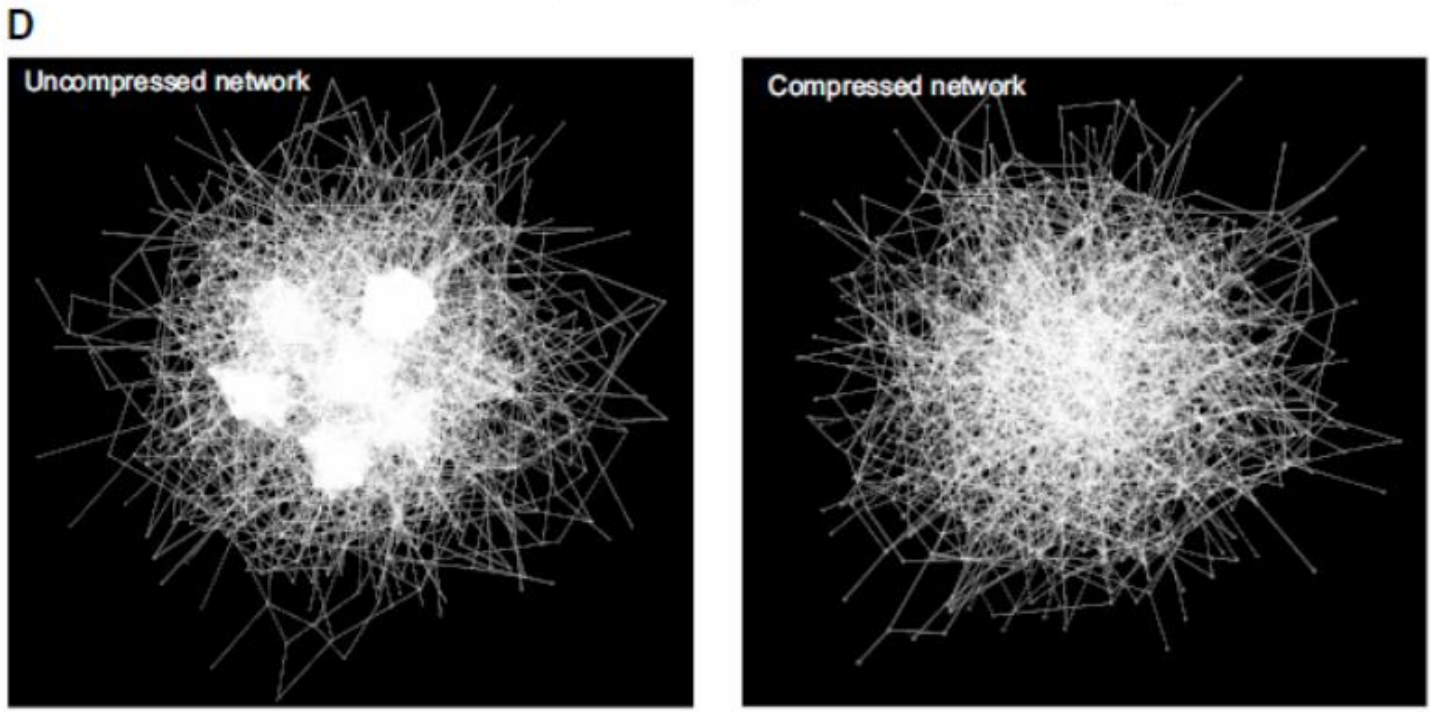
curacy values for the test set of 1,000 fully annotated (EC numbers) enzyme sequences.

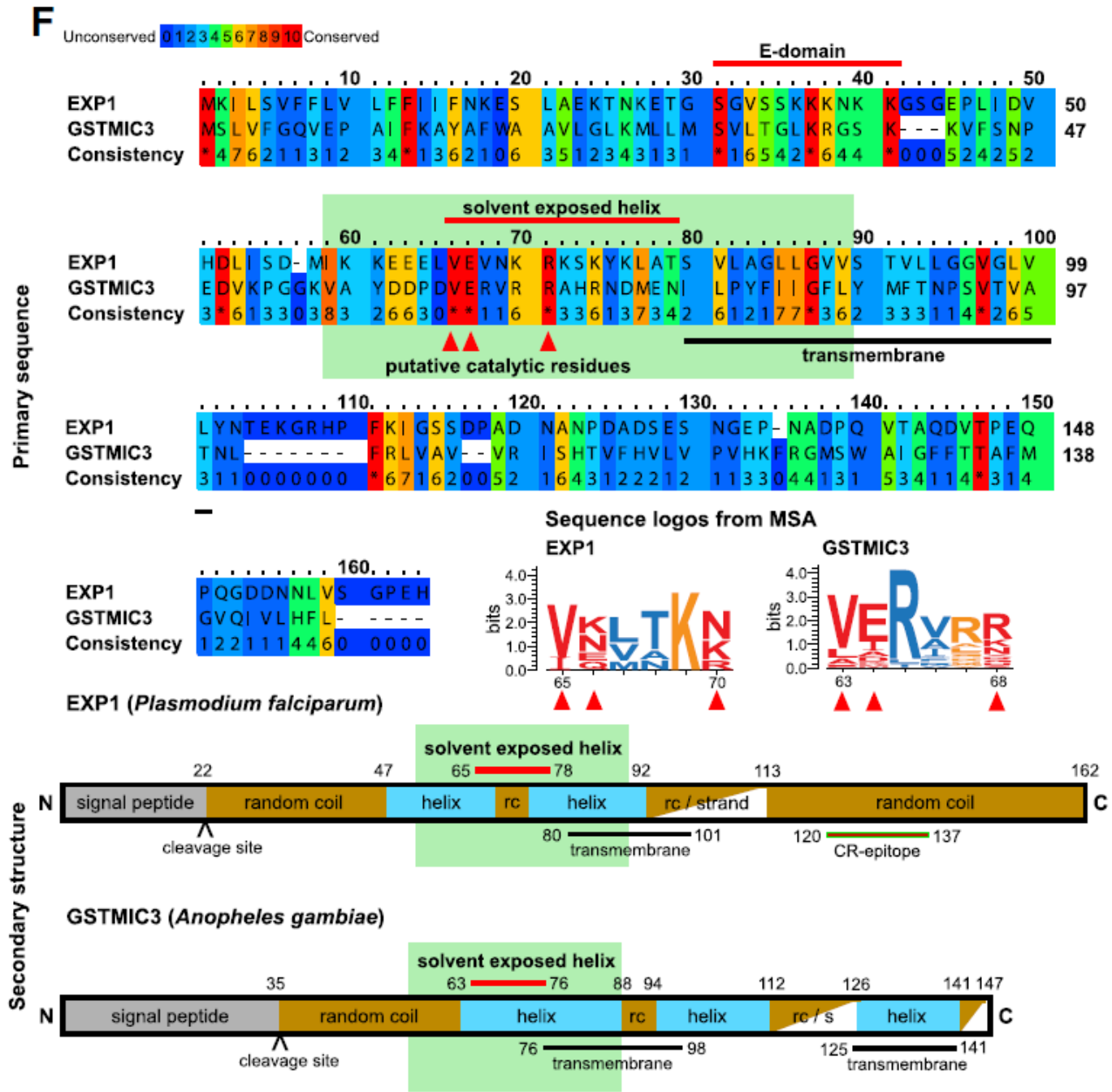
s RankProp (G) scatter plots for the same test set.

or sequence similarity net-
l-BLAST sequence search
rom the full uncompressed
and the ranking of network
similarity to the query pro-
method takes into account
ted to intrinsic evolutionary
1,000 enzymes from the
UniProt database (UniProt
et of the leave-one-out ex-
we measured success by
predictions made prior to reaching 50 false predictions, or ACC_{50} accuracy values (Gribskov and Robinson, 1996). We found that GID Z scores above $Z \sim 2$ were reliable indicators of functional accuracy (Figure 2D) giving accuracies that were 9% greater for GID than for PSI-BLAST ($p < 9.2 \times 10^{-17}$, Wilcoxon signed rank test; Figure 2E) and 21% more than with COGs ($p < 1.7 \times 10^{-11}$; Figure 2F). Our results showed that local algorithms on the full, uncompressed network (such as PSI-BLAST) yielded lower prediction accuracies than GID on the compressed network.

We also benchmarked GID against an established global algorithm, RankProp (Melvin et al., 2009). The iterative flow algorithm underlying RankProp was computationally intractable on the


A
Uncompressed network
B
Intra-COG compression
C
Intra- and inter-COG compression
Intrinsic link
Contextual link
Protein node
Single proteome
COG
New COG core node


D
Uncompressed network
Compressed network


F
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved
E-domain
EXP1
GSTMIC3
Consistency
50
47
solvent exposed helix
putative catalytic residues
transmembrane
99
97
148
138
Primary sequence
Sequence logos from MSA
EXP1
GSTMIC3
bits
160
EXP1 (Plasmodium falciparum)
signal peptide
random coil
helix
rc
rc / strand
cleavage site
CR-epitope
GSTMIC3 (Anopheles gambiae)
Secondary structure

BCB 5200 Introduction to Bioinformatics

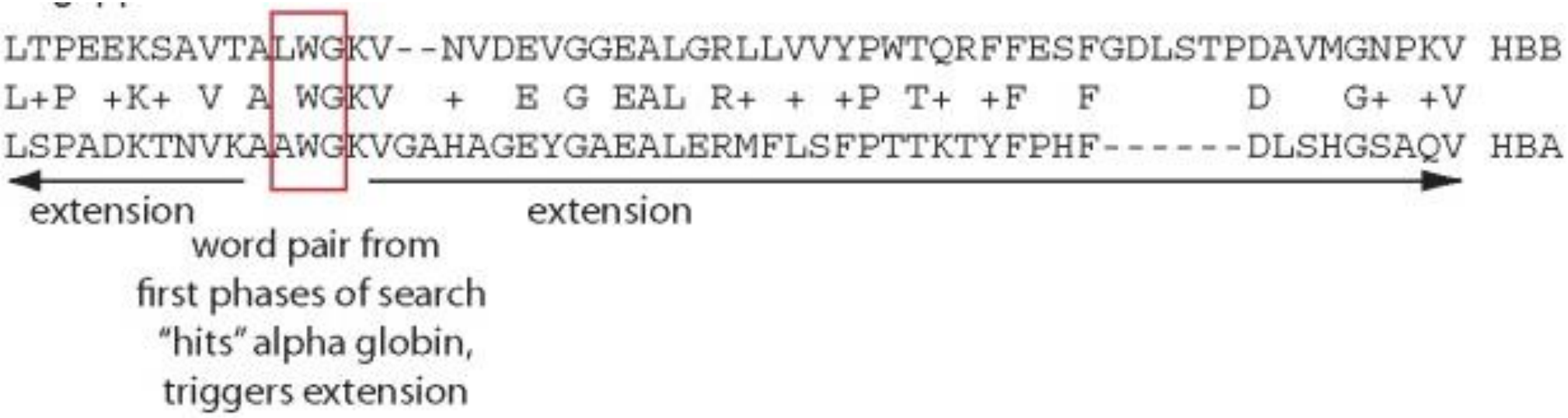
Profile sequence searches

Bioinformatics and Computational Biology
Saint Louis University

How a BLAST search works

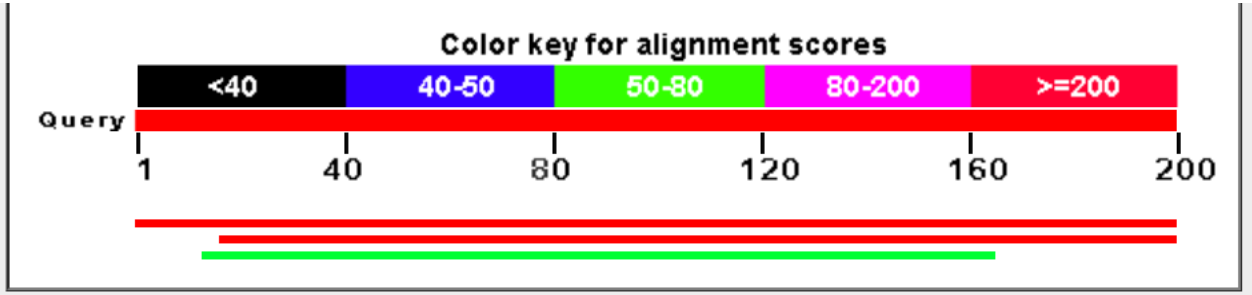
"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)



Limitation of BLAST

- To find other human homologs of human retinol-binding protein 4 (NP_006735)
- BLASTP (refseq+BLOSUM62)



Descriptions

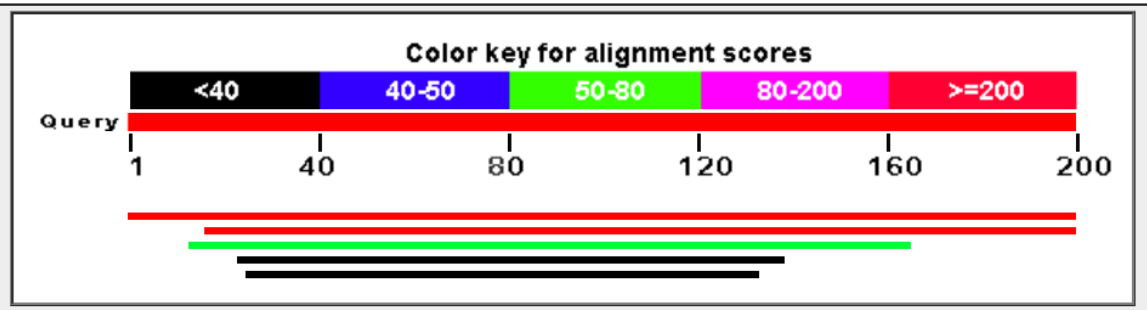
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Ident	Links
NP_006735.2	retinol-binding protein 4 isoform a precursor [Homo sapiens]	420	420	100%	4e-149	100%	GMP
NP_001310447.1	retinol-binding protein 4 isoform b [Homo sapiens]	386	386	92%	8e-136	99%	
NP_001638.1	apolipoprotein D precursor [Homo sapiens]	55.5	55.5	76%	3e-06	28%	GM

Limitation of BLAST

- To find other human homologs of human retinol-binding protein 4 (NP_006735)
- BLASTP (refseq+BLOSUM45)



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Ident	Links
NP_006735.2	retinol-binding protein 4 isoform a precursor [Homo sapiens]	438	438	100%	3e-160	100%	GM
NP_001310447.1	retinol-binding protein 4 isoform b [Homo sapiens]	404	404	92%	1e-146	99%	
NP_001638.1	apolipoprotein D precursor [Homo sapiens]	51.2	51.2	76%	7e-08	28%	GM
NP_000597.2	complement component C8 gamma chain precursor [Homo sap	32.7	32.7	57%	0.26	25%	GM
NP_001624.1	protein AMBP preproprotein [Homo sapiens]	32.4	32.4	54%	0.48	23%	GM

“Twilight zone” of protein homologs

- Similarity sequence searches (e.g. BLAST, FASTA) are known to miss the hits with 10%-20% of similarity. This “area” of similarity is called the “Twilight zone”.

- Some protein families might have fast evolving rate or highly divergent and so it will be hard to
  1) identify complete orthologs and paralogs
  2) determine whether any homolog is present in a particular organism

- Identifying distant relationship between protein subfamilies, families, or even superfamilies.

Outline

- Protein evolution and conservation patterns
  - Motifs/Patterns
  - Profiles

- Position-specific scoring matrix (PSSM)
  - PSI-BLAST
  - RPS-BLAST
  - DELTA-BLAST
  - PHI-BLAST

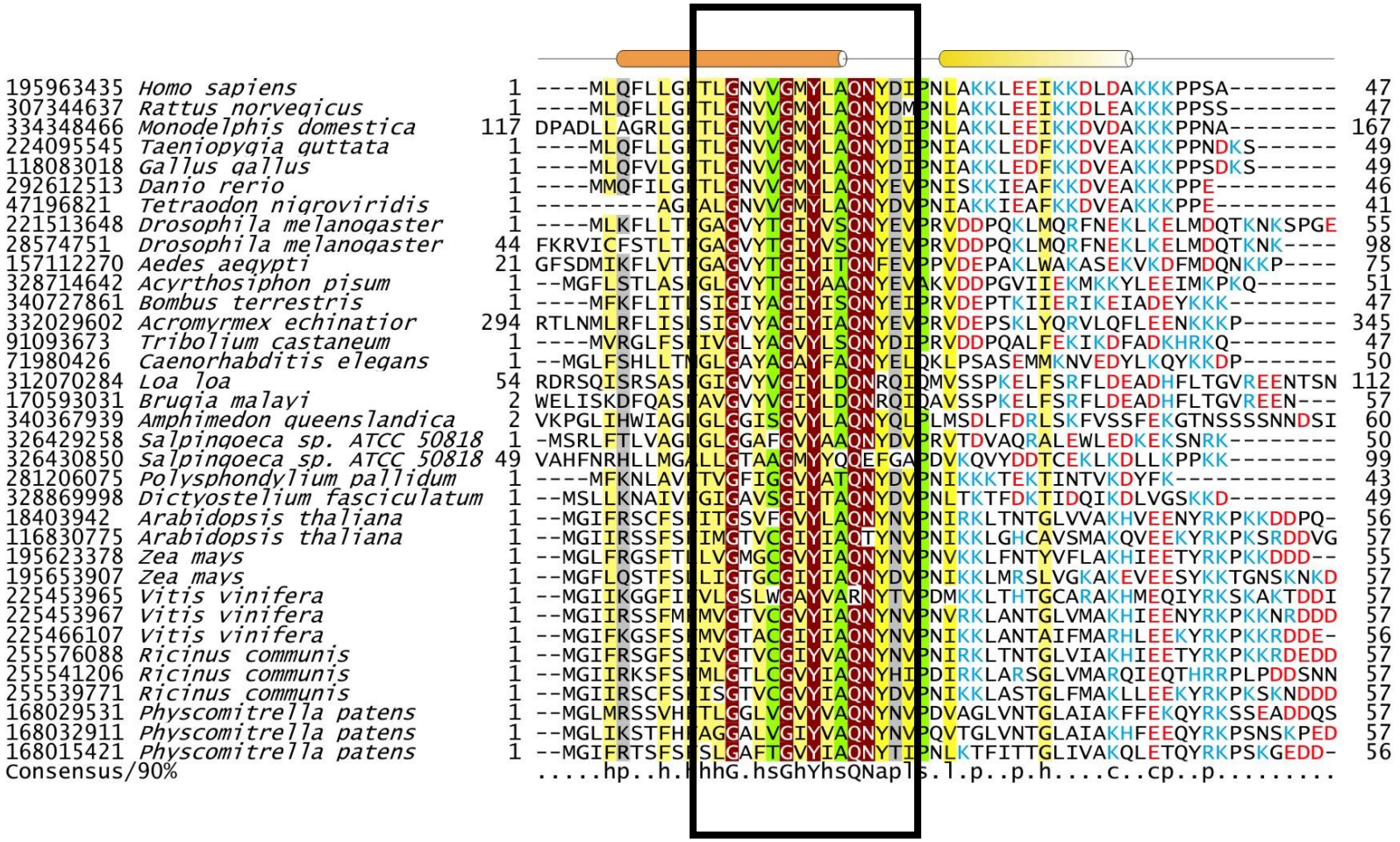
- Hidden markov models
  - HMMer

Protein evolution and conservation patterns

- In the 1990's people began to see that aligning sequences to profiles (multiple sequence alignment) gave much more information than pairwise alignment alone.

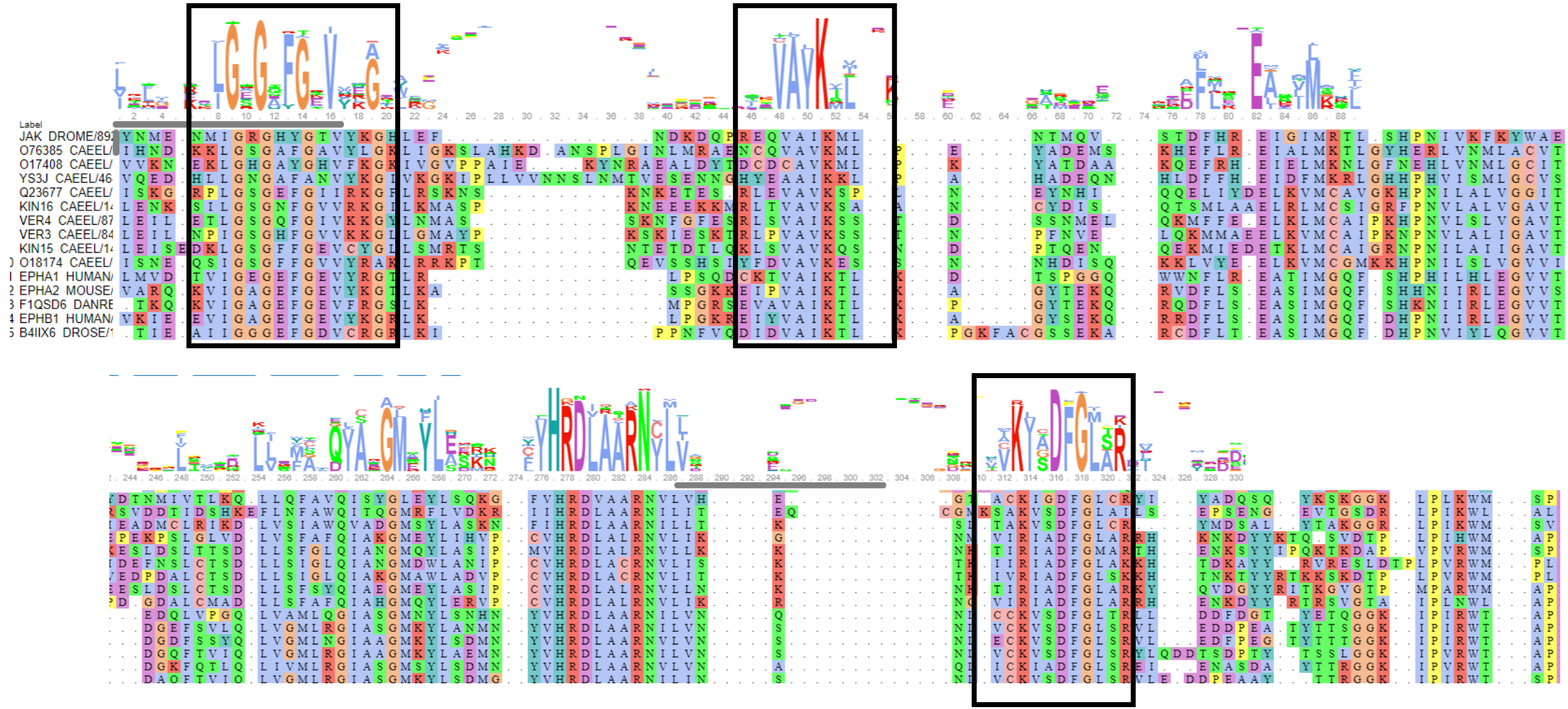
- During evolution, protein sequences diverge on many positions whereas retain conservation on some positions due to structural or functional constraints. Those characteristic residues are called as protein motifs or patterns:
  - Active site or catalytic site
  - Ligand binding pocket
  - Structural core

Protein motifs/patterns: TM protein

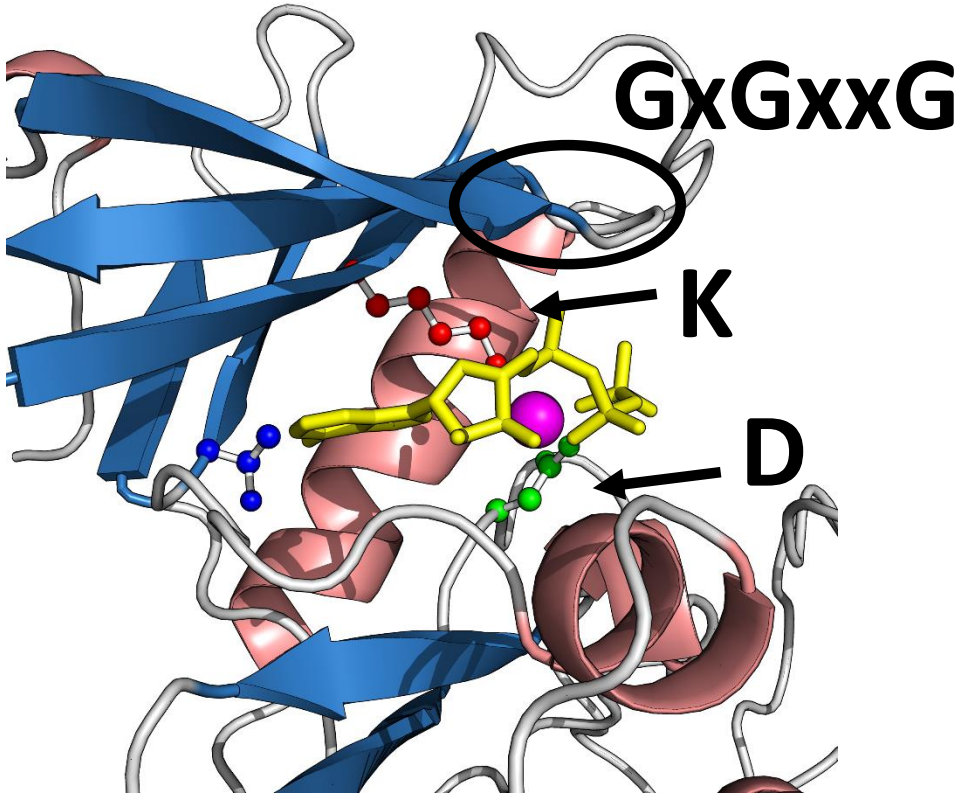


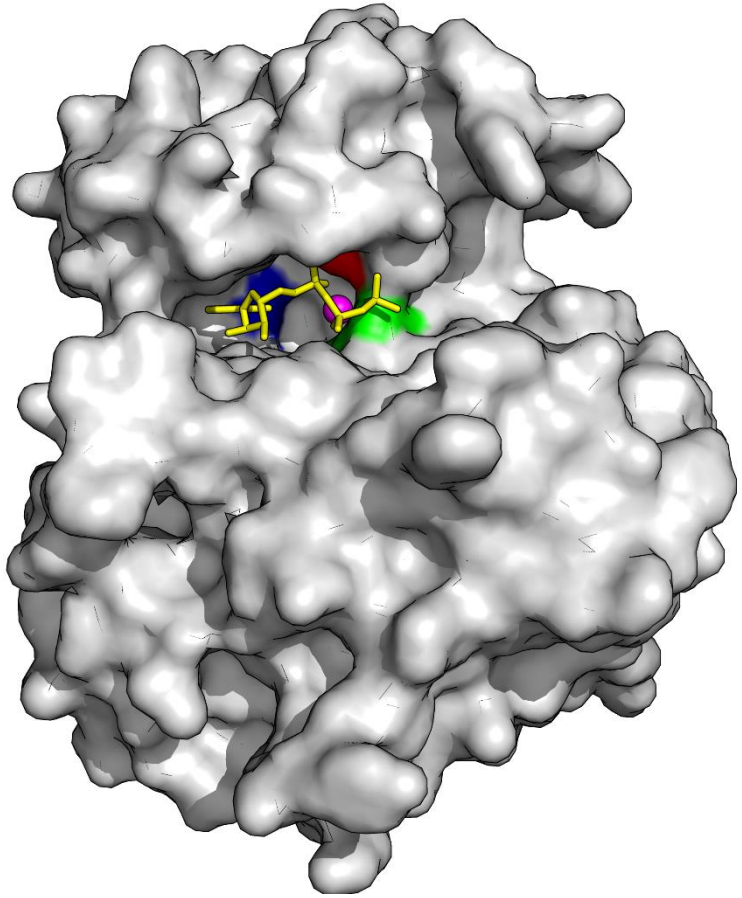
Consensus: GxxxGxYxxQN

Protein motifs/patterns: protein kinase

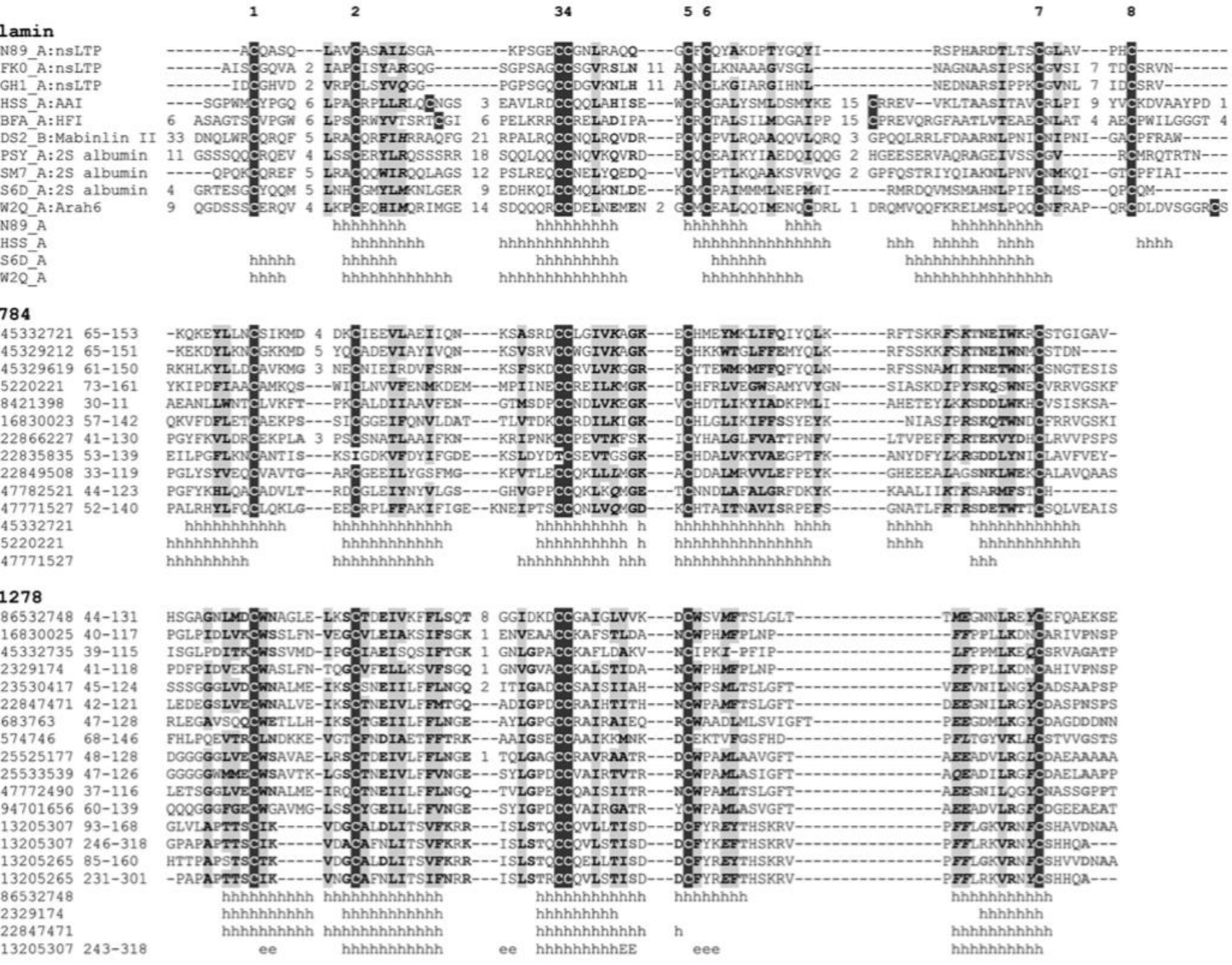


Protein motifs/patterns: protein kinase

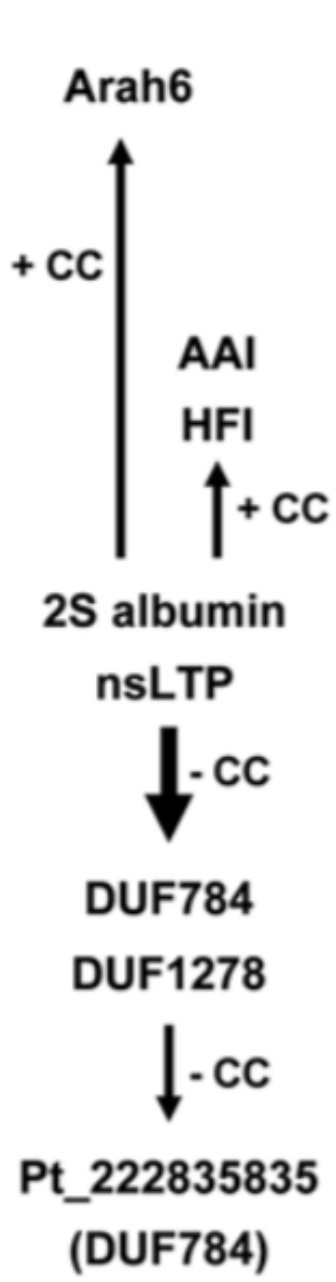


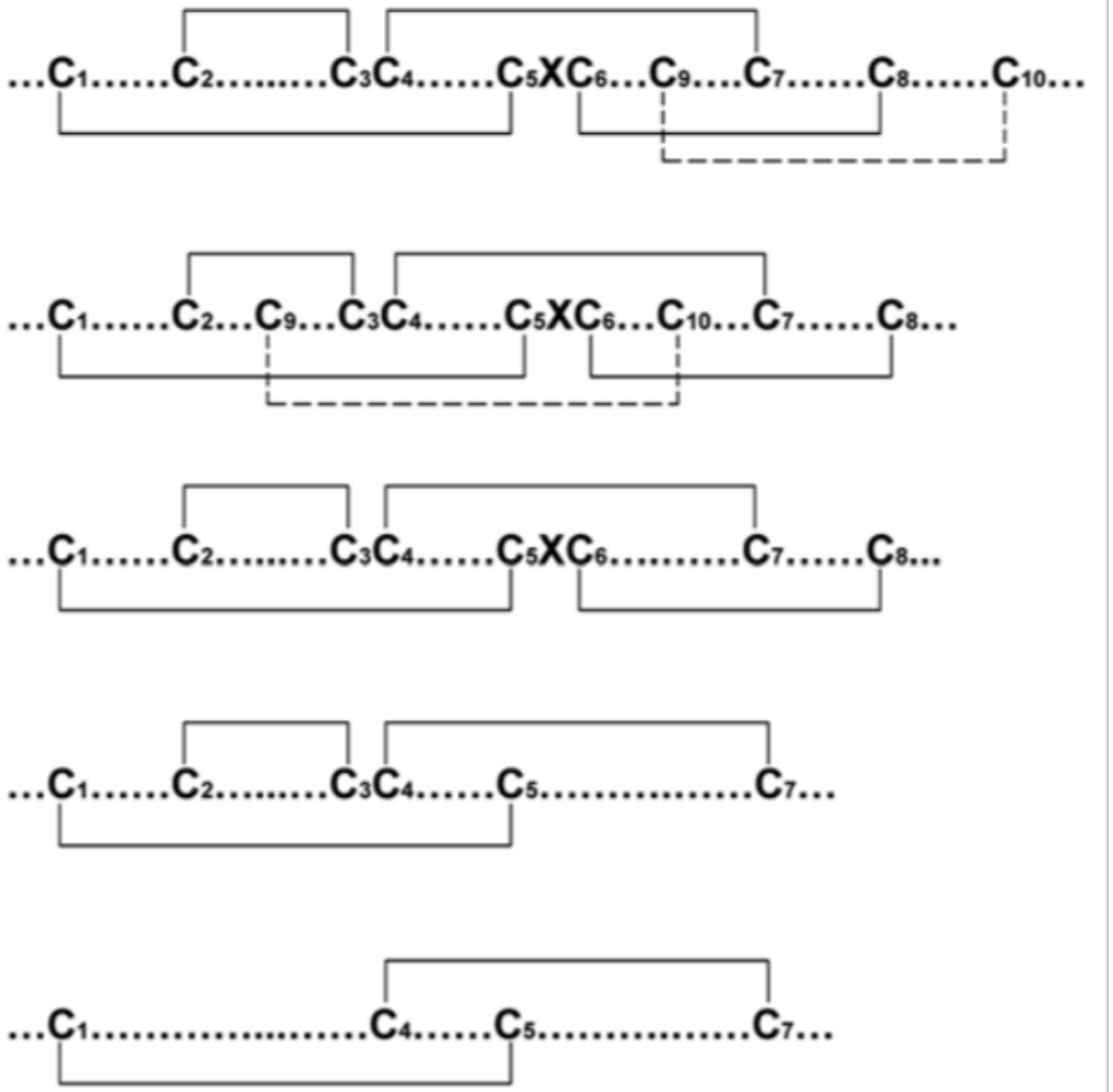


Plant lipid-transfer protein families



Plant lipid-transfer proteins





Profile: protein position specific information

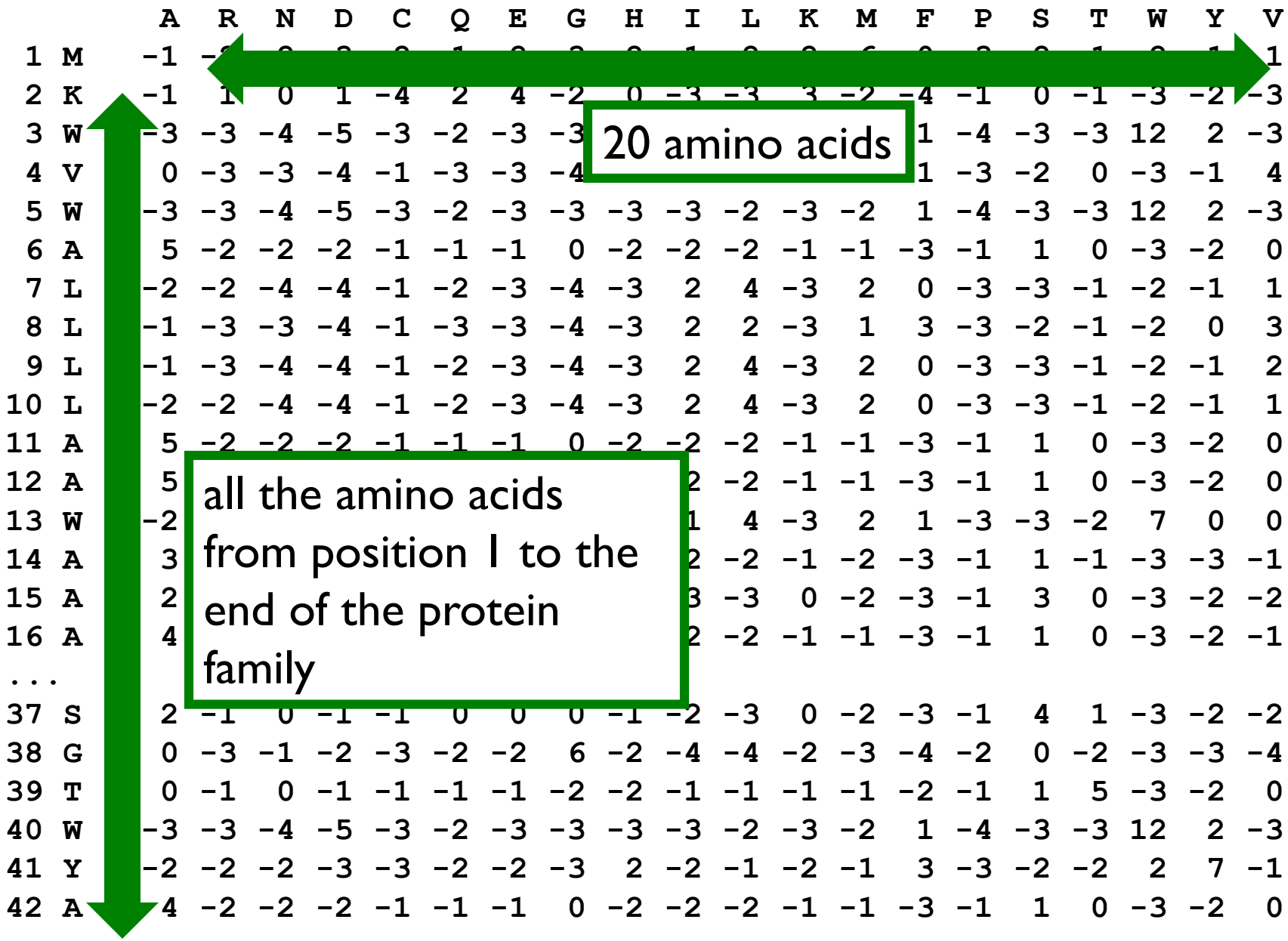
- For protein families, some positions are more important (less likely to change) than others
  - Better at characterizing structural/functional properties of proteins
  - More sensitive at identifying distant homologies

- **Profile:** A table that lists the frequencies of each amino acid in each position of protein sequence. Frequencies are calculated from a MSA containing a domain of interest
  - Allows us to identify consensus sequence
  - align a new sequence to the profile
  - Profile can be used in database searches

- Position-specific scoring matrix (PSSM) to model amino acid substitutions on each position.

Position-specific scoring matrix

- A PSSM is an n by m matrix, where n is the size of the alphabet, and m is the length of the sequence.



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	0	2	0	1	0	2	0	1	0	0	6	0	2	0	1	0	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	1	-2	1	-4	-3	-3	12	2	-3	
4	V	0	-3	-3	-4	-1	-3	-3	-4					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	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2									1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3									2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2									3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4									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

Position-specific scoring matrix

- A PSSM is an n by m matrix, where n is the size of the alphabet, and m is the length of the sequence.
- The entry at (i, j) is the score assigned by the PSSM to letter i at the jth position.

		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	-2	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	4	-3	2	0	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-2	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-2	-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

Position-specific scoring matrix

		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									2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4									2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2									1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2									1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4									2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2									2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1											2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2									1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1									2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2									3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-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 current protein family can receive different scores for matching alanine—depending on the position in the protein

Position-specific scoring matrix

		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 current protein family can receive different scores for matching tryptophan—depending on the position in the protein

Position-specific scoring matrix

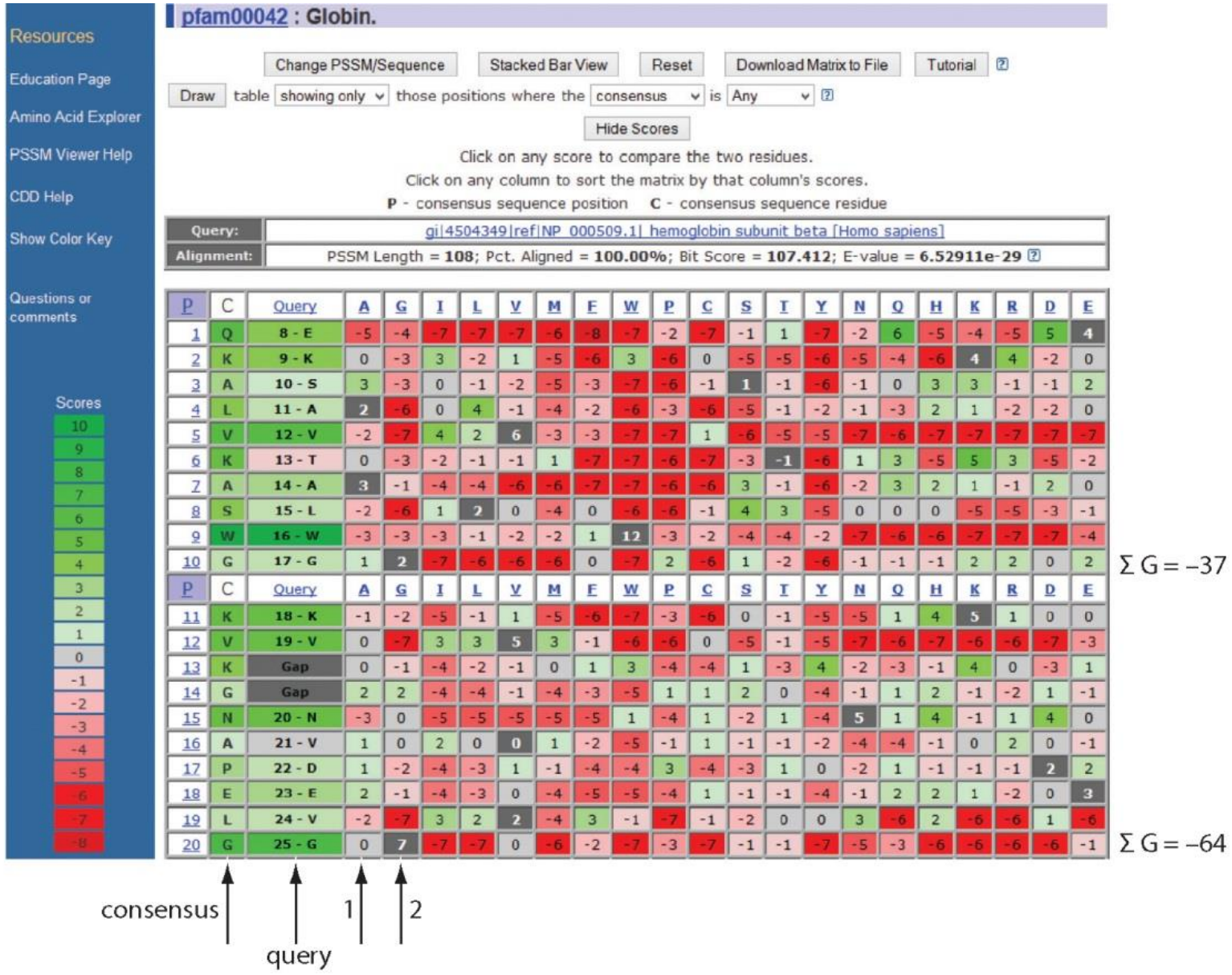
- This PSSM assigns the sequence a score at:

M E F W Y A G I

6 +4 +1 -3+2+5-4+2 = 13

		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

Position-specific scoring matrix



Negative scores indicate that alignment of a Gly with any other residue occur less frequently than expected by chance.

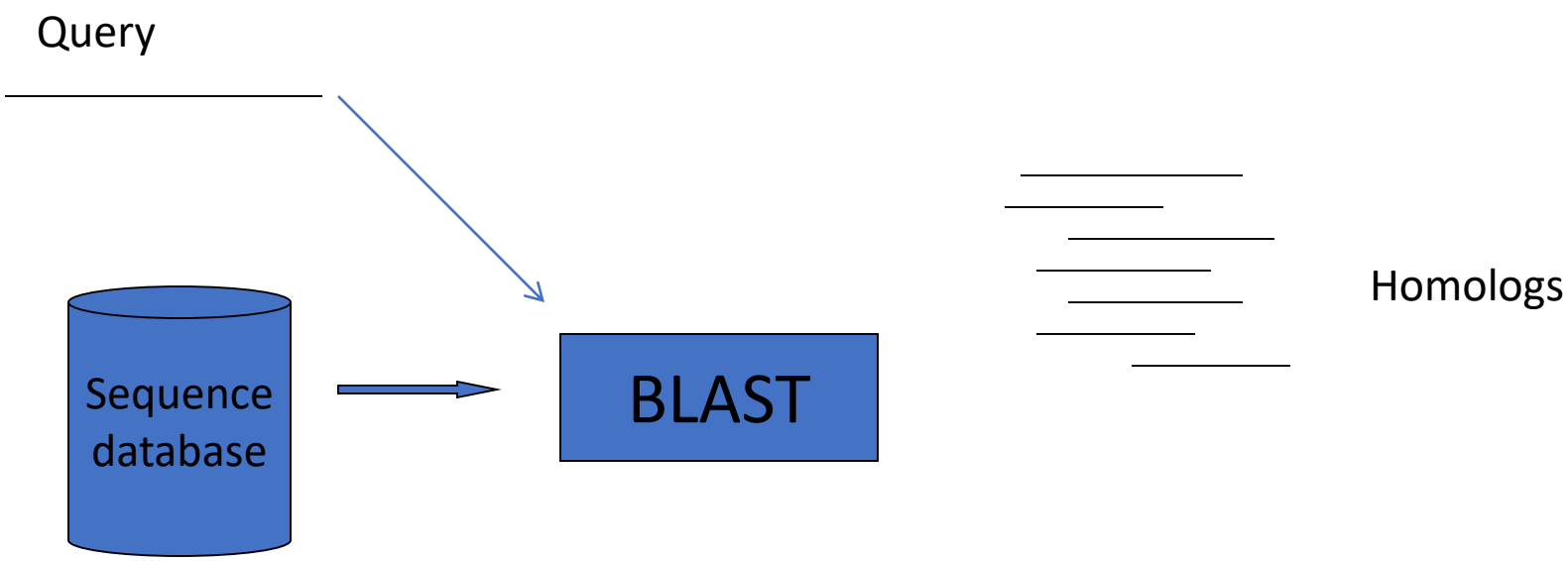
Occurrence of Gly is heavily favored and the penalties for not selecting a Gly are severe.

The PSSM reflects a more customized estimate of amino acid substitutions at all positions

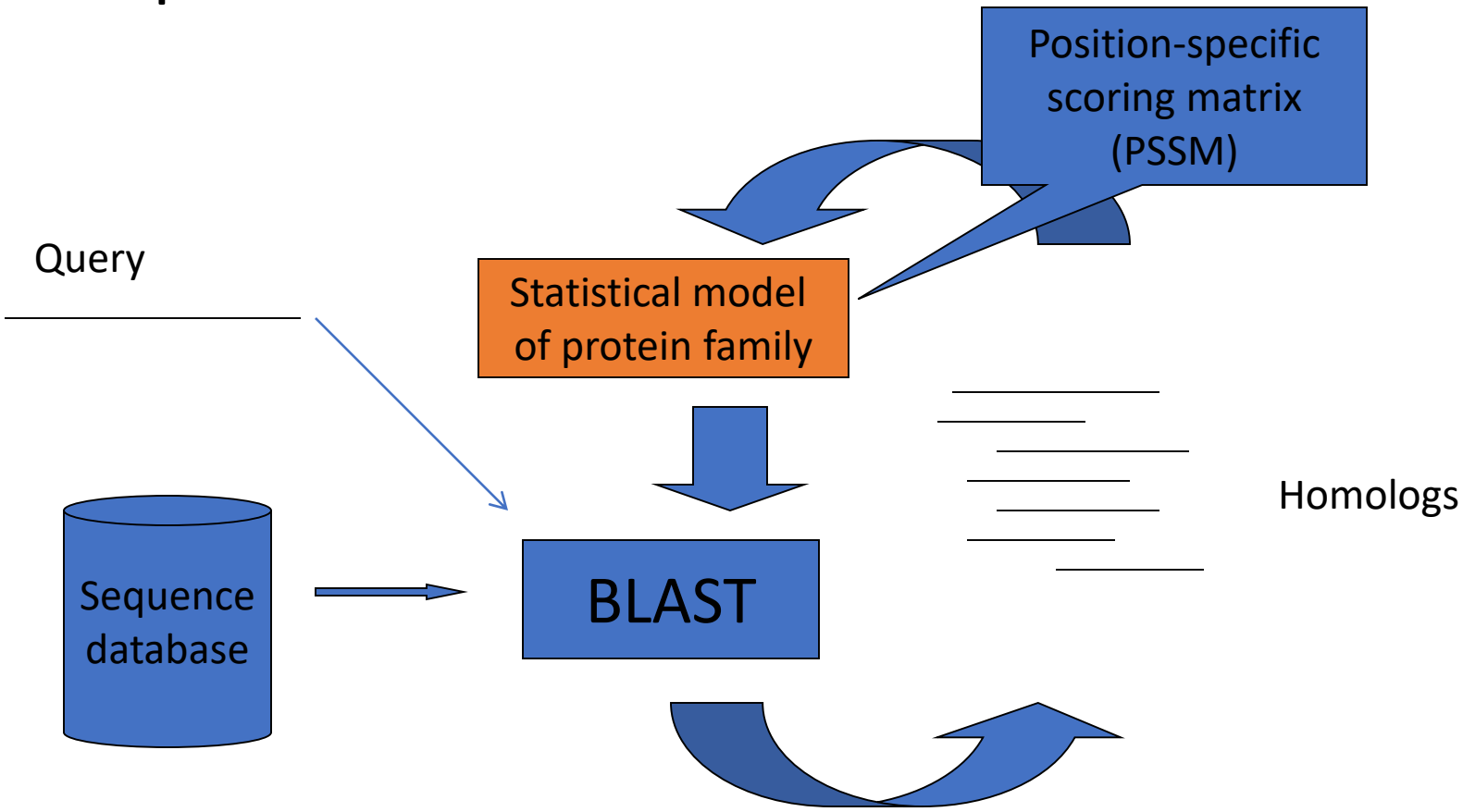
Position specific iterated BLAST: PSI-BLAST

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

BLAST



Position-specific iterated BLAST



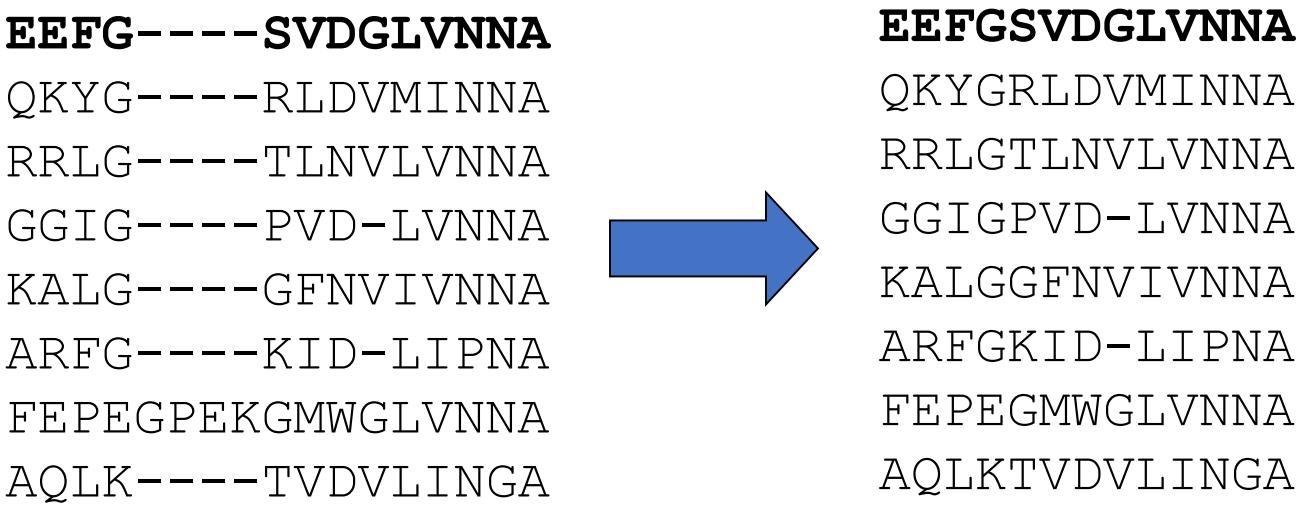
PSI-BLAST is performed in five steps

[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)

Creating a PSSM from multiple sequences

- Discard columns that contain gaps in the query.



Creating a PSSM from multiple sequences

- Discard columns that contain gaps in the query.

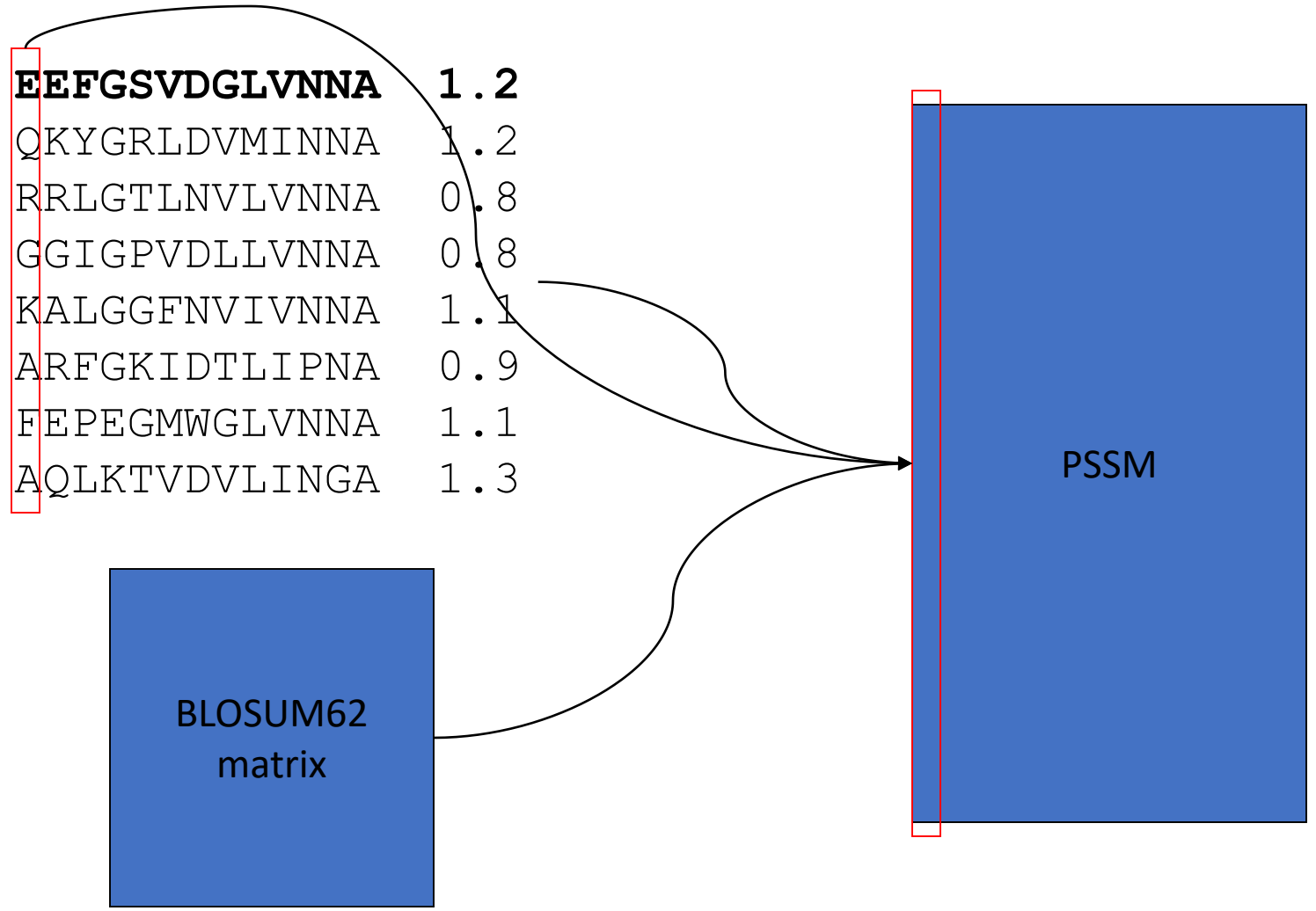
- For each column C
  - Compute relative sequence weights
  - Compute PSSM entries, taking into account
    - Observed residues in this column
    - Sequence weights
    - Substitution matrix

Compute sequence weights

EEFGSVDGLVNNA	**1.2**
QKYGRLDVMINNA	1.2
RRLGTLNVLVNNA	0.8
GGIGPVDLLVNNA	0.8
KALGGFNVIVNNA	1.1
ARFGKIDTLIPNA	0.9
FEPEGMWGLVNNA	1.1
AQLKTVDVLINGA	1.3

- Low weights are assigned to redundant sequences.
- High weights are assigned to unique sequences.

Compute PSSM entries



Creating a PSSM from multiple sequences

- The score the profile for amino acid a at position p is
- $$M(p,a) = \sum_{b=1}^{20} f(p,b) \cdot s(a,b)$$

  where
  - f(p,b) = frequency of amino acid b in position p
  - s(a,b) is the score of (a,b) (from, e.g., BLOSUM or PAM)

PSI-BLAST is performed in five steps

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

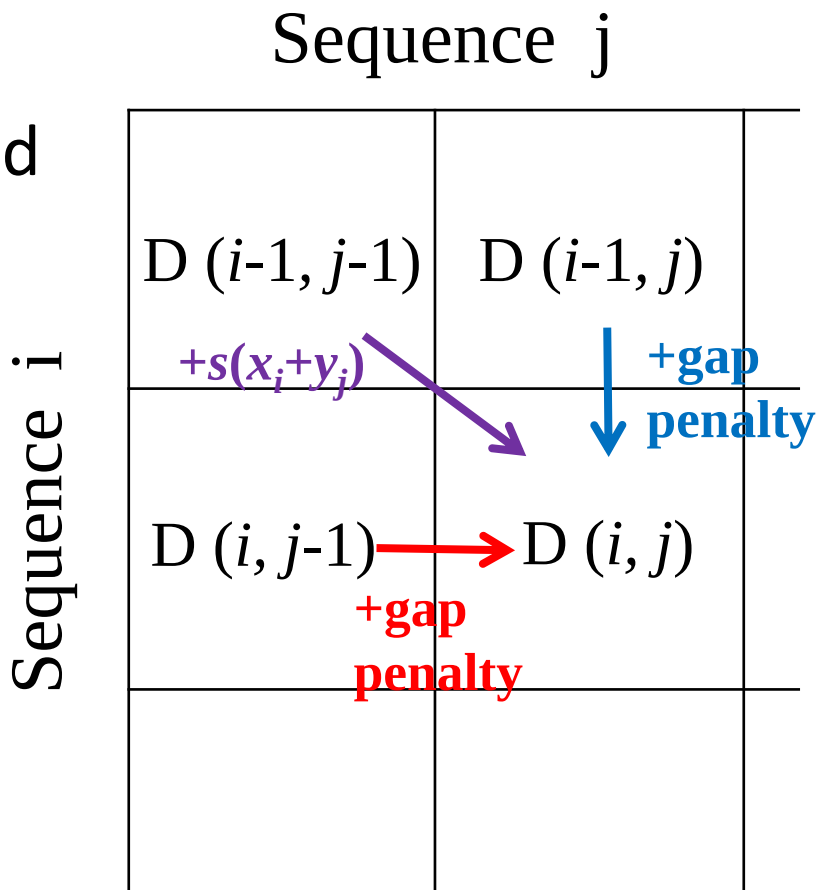
[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

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

Profile-to-sequence alignments

- Optimal alignment can be found by dynamic programming
  - Extension of Needleman-Wunsch
- Spaces are only added to msa – never removed
  - Once a gap, always a gap

$$D(i,j) = \max \begin{cases} D(i-1,j-1) + s(x_i, y_j) \\ D(i-1,j) + g \\ D(i,j-1) + g \end{cases}$$



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) and more sequences are identified

	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) and more sequences are identified

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

PSI-BLAST: dramatic increase in number of hits

Human beta globin (NP_000509) was used as a query in a PSI-BLAST search of the RefSeq database restricted to fungi.

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

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

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

(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
```

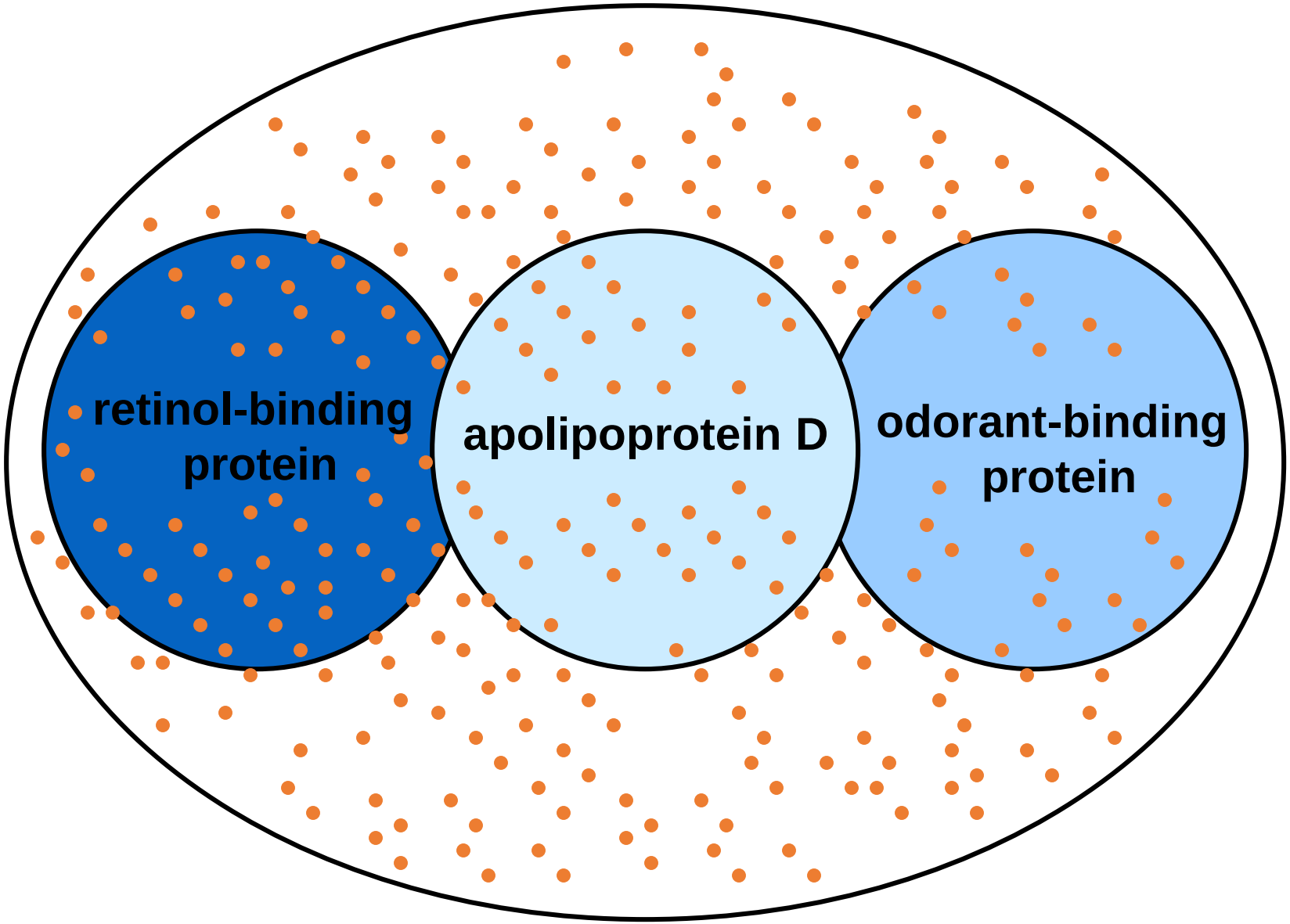
(c) PSI-BLAST iteration 3 (human beta globin versus a *C. albicans* globin)

Range 1: 281 to 426 GenPept Graphics

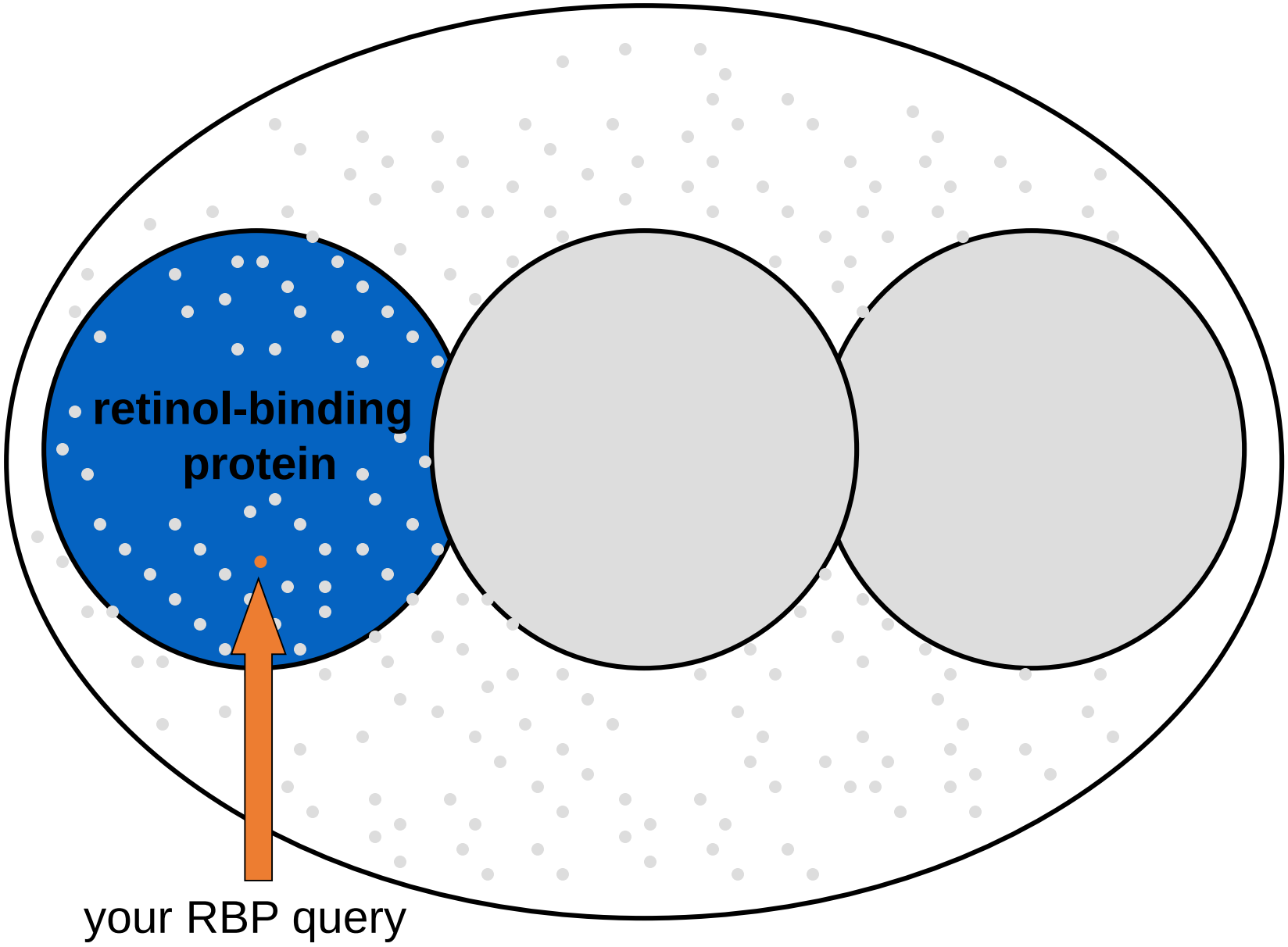
Score	Expect	Method	Identities	Positives	Gaps
128 bits(321)	2e-33	Composition-based stats.	28/146(19%)	50/146(34%)	6/146(4%)

```
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
```

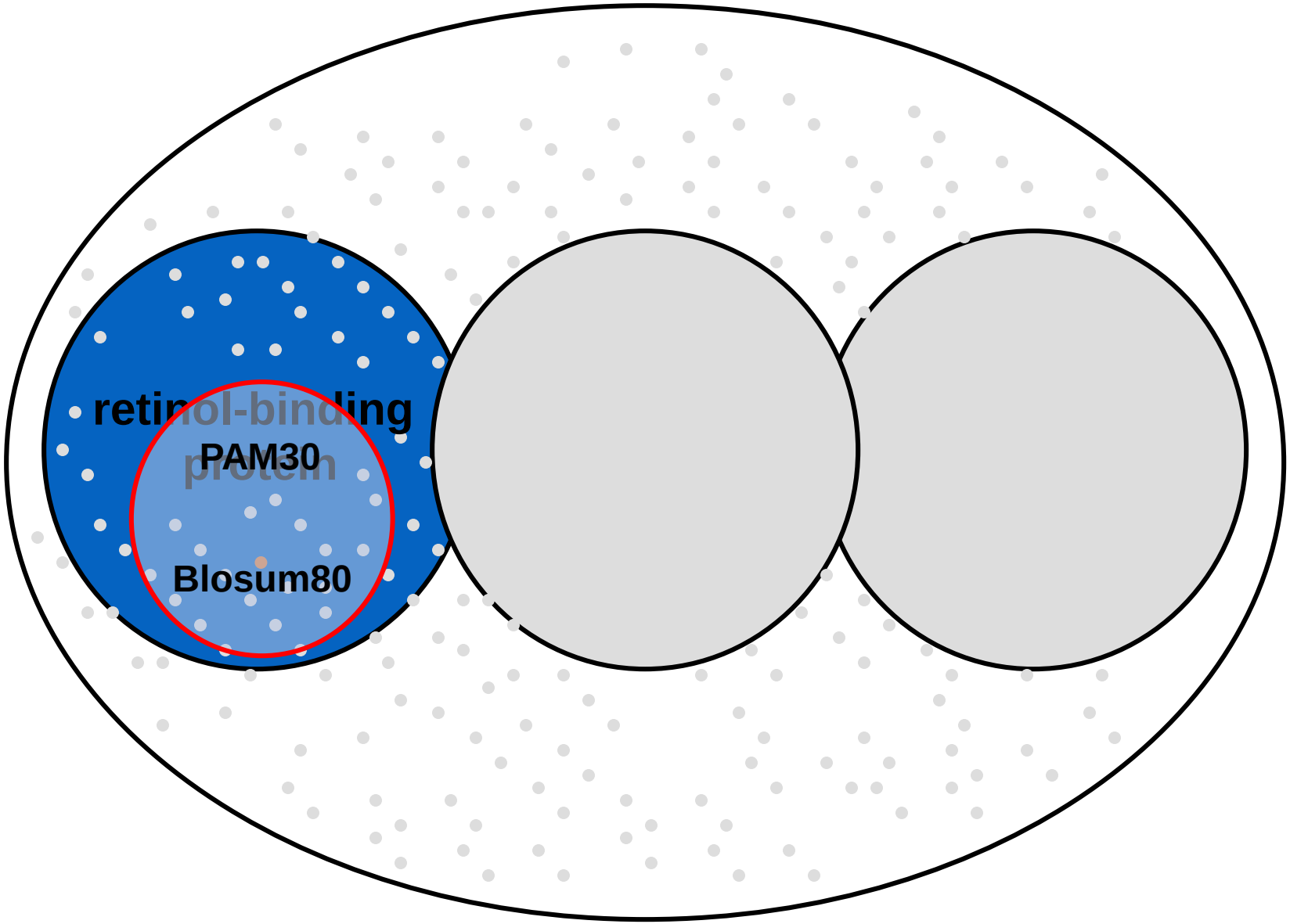
The universe of lipocalins (each dot is a protein)



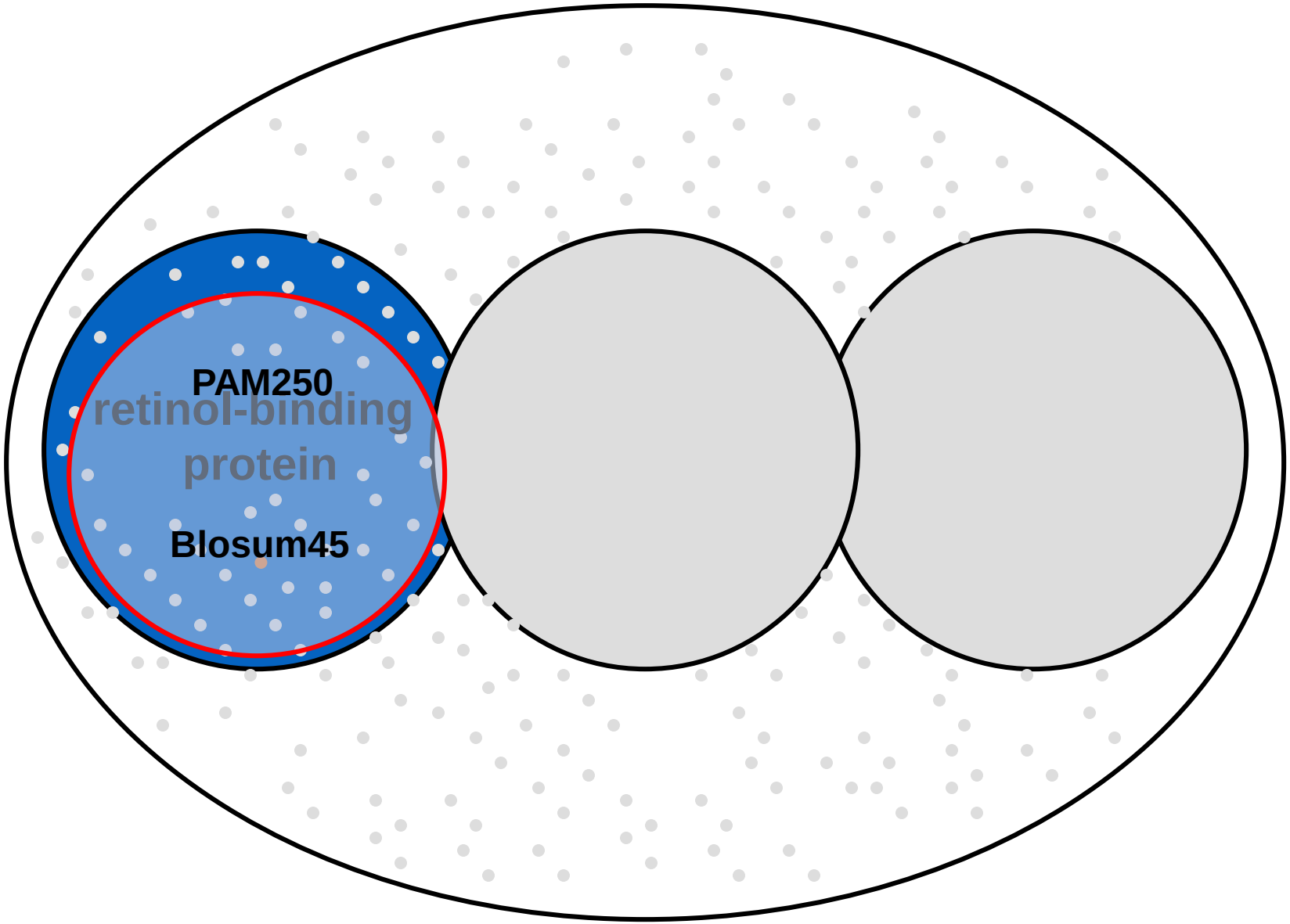
Scoring matrices let you focus on the big (or small) picture



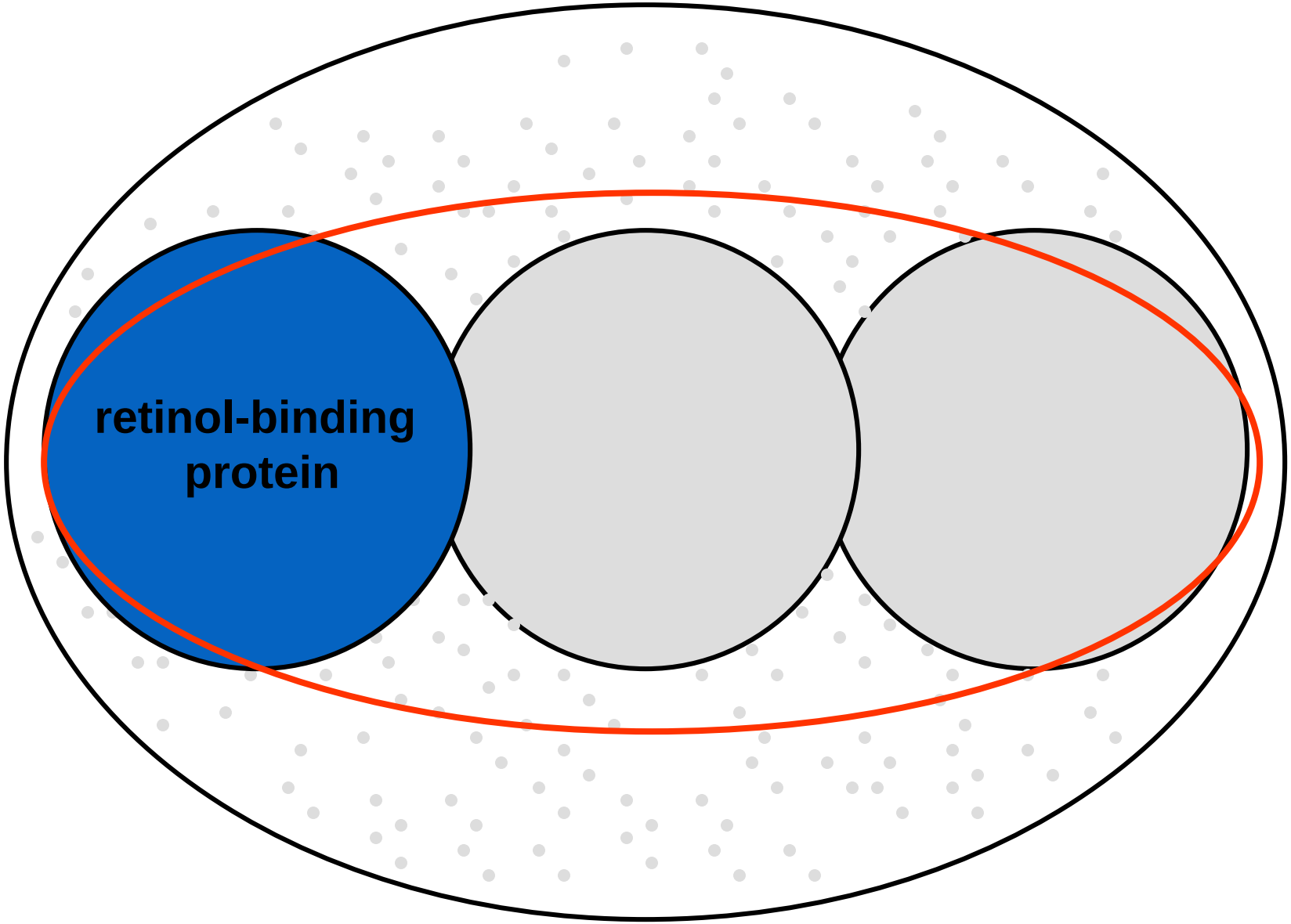
Scoring matrices let you focus on the big (or small) picture



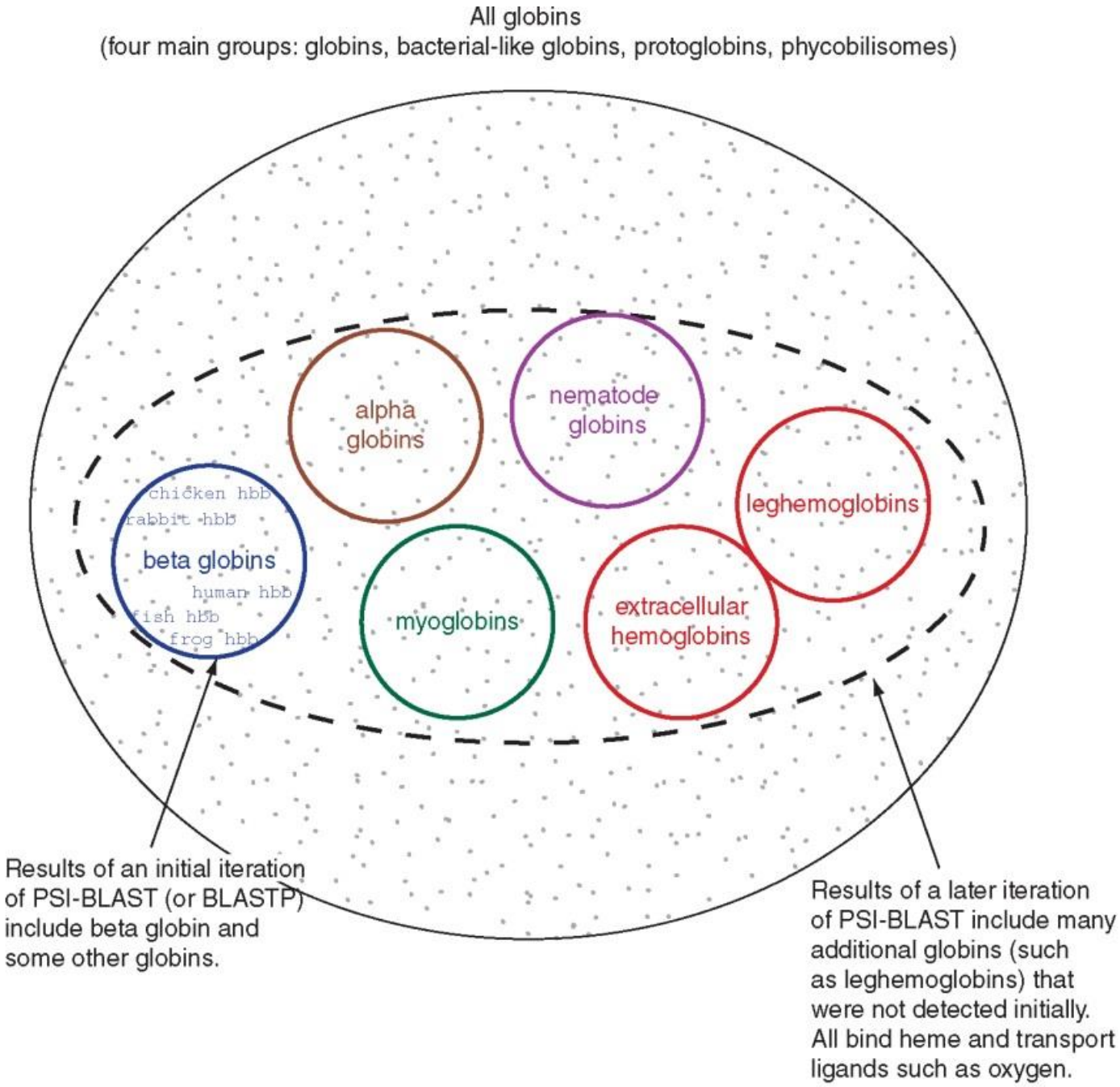
Scoring matrices let you focus on the big (or small) picture



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, a match, even if it is a false positive (that is not truly homologous to the query), can be incorporated into a PSSM. This will lead to the inclusion of many other related false positive hits in later iterations.

There are three main approaches to removing false positives:

(1) Filter biased amino acid regions. (Automatic or manually)
(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).

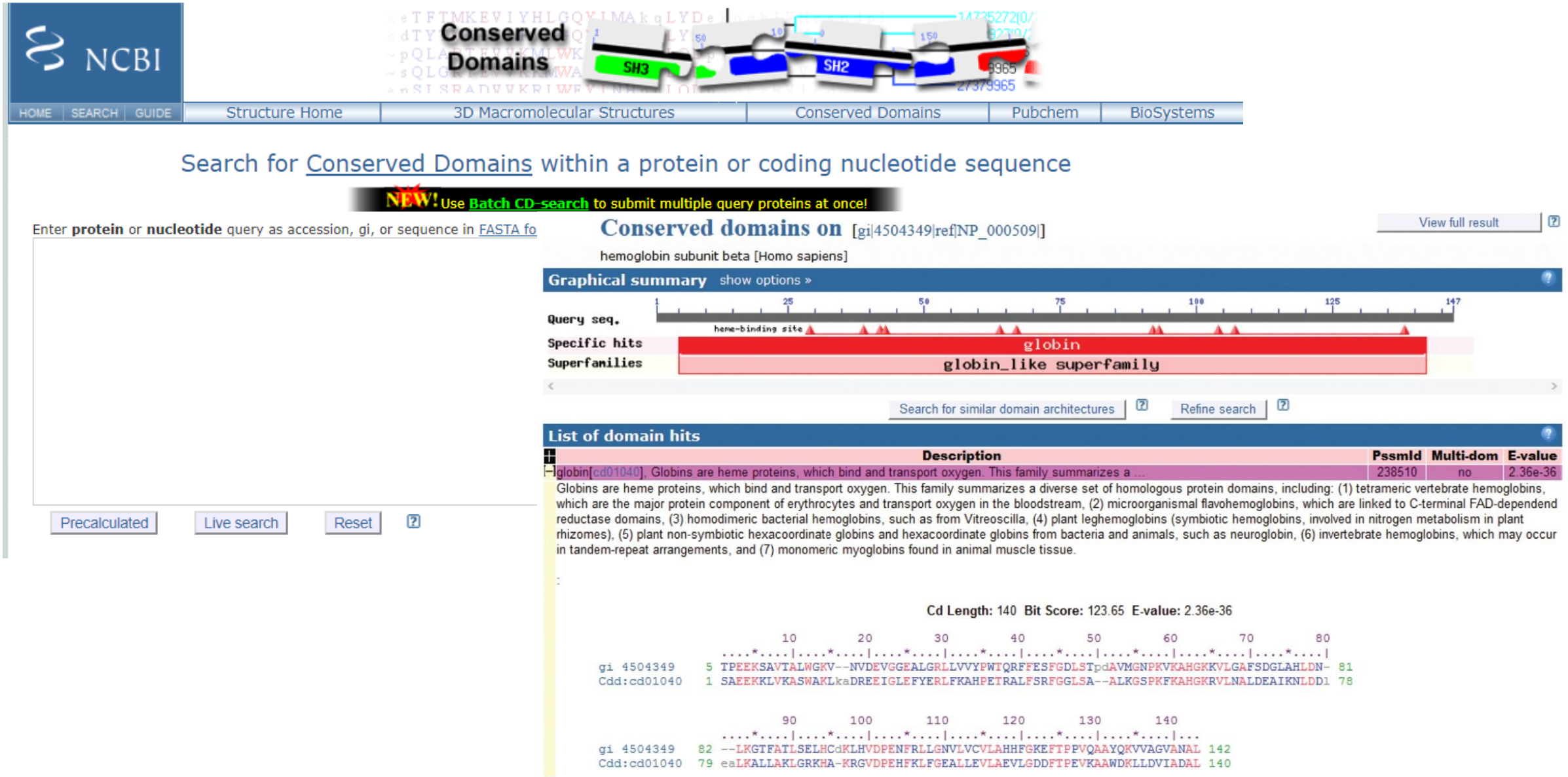
Conserved Domain Database (CDD)

- Profiles or PSSMs can be used as a feature of protein domains or families.

- NCBI-CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models of domains and their position-specific score matrices (PSSMs).

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

RPS-BLAST searches are incorporated into the NCBI-CDD

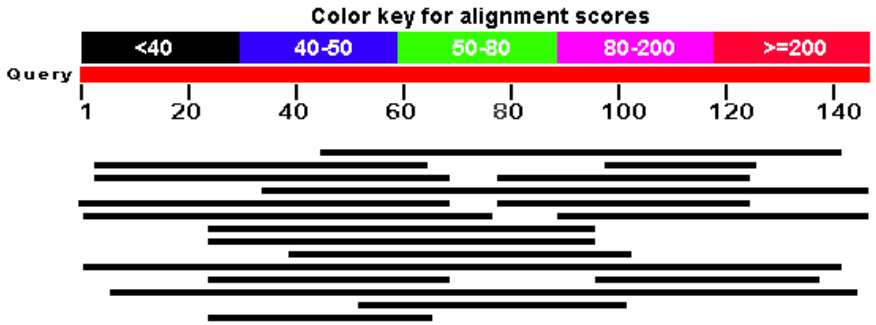


Domain enhanced lookup time Accelerated BLAST (DELTA-BLAST)

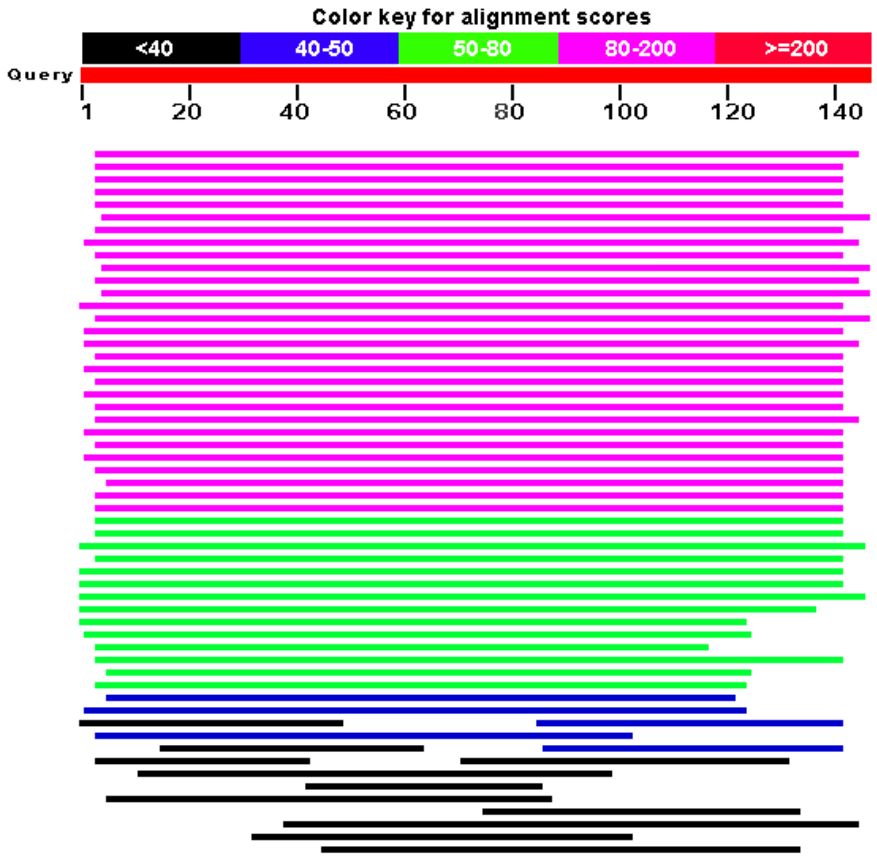
- DELTA-BLAST searches a query against a library of pre-computed PSSMs and use it for BLAST search.
  - Most queries do match a PSSM; if not the search proceeds in a PSI-BLAST-like manner.
  - One iteration of DELTA-BLAST is recommended.

Search HBB (NP_000509) against RefSeq plants…

BLAST



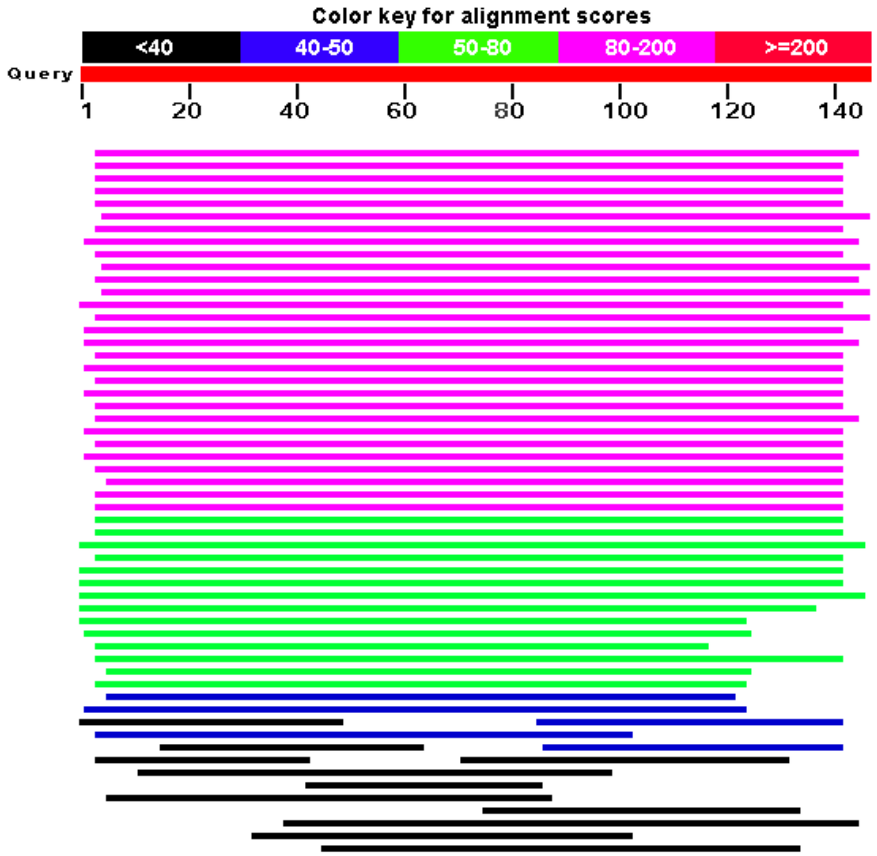
DELTA-BLAST



DELTA-BLAST

DELTA-BLAST might be better than PSI-BLAST given it takes advantage of longer PSSMs (quality varies)

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



Pattern hit initiated BLAST (PHI-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).

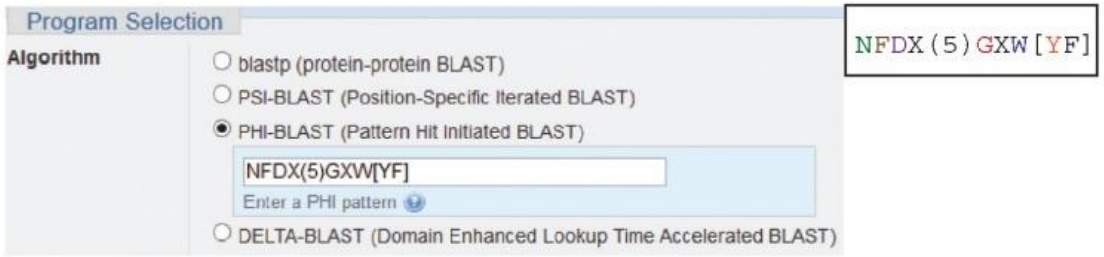
PHI-BLAST is specialized, and is not commonly used but can be very useful.

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



(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.

Summary

- Profiles and PSSMs
- PSI-BLAST creates multiple alignments and position-specific scoring matrices (PSSMs) which are used to search database.
- RPS-BLAST searches a query against a collection of predefined position-specific scoring matrices
- DELTA-BLAST searches a query against a library of pre-computed PSSMs and use it for BLAST search.

Practice: Exp1 and GSTMIC3

EXP1: circumsporozoite-related antigen (AAA21753; Plasmodium falciparum)

GSTMIC3: microsomal glutathione transferase (AAP37005, Anopheles gambiae)

Try PSI-BLAST in plasmodium, Anopheles and NR, check the sequence features