CHEM 436 / 630

Molecular modelling of proteins

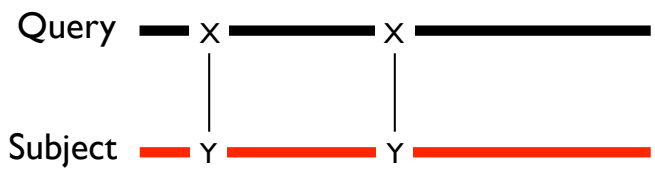
Winter 2018 Term

Instructor:
Guillaume Lamoureux
Concordia University, Montréal, Canada

BLAST *versus* PSI-BLAST

BLAST uses scoring matrices (PAM or BLOSUM) that are independent of the position of the amino acid in the sequence.

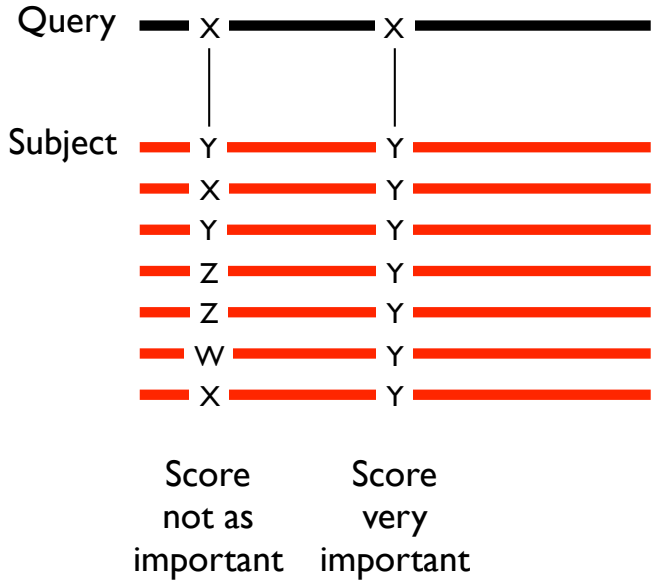
This is arguably the best we can do if we don't know what are the important domains of the query sequence.



The same X-Y pairing contributes the same to the total score wherever it is in the sequence.

However, once we have performed a BLAST search and found a large number of matches, we have a better idea of what the important residues are and we can have the score of these residues carry more weight.

This information can be encoded into a Position-Specific Scoring Matrix (PSSM), for which each position along the sequence has its own scoring table.



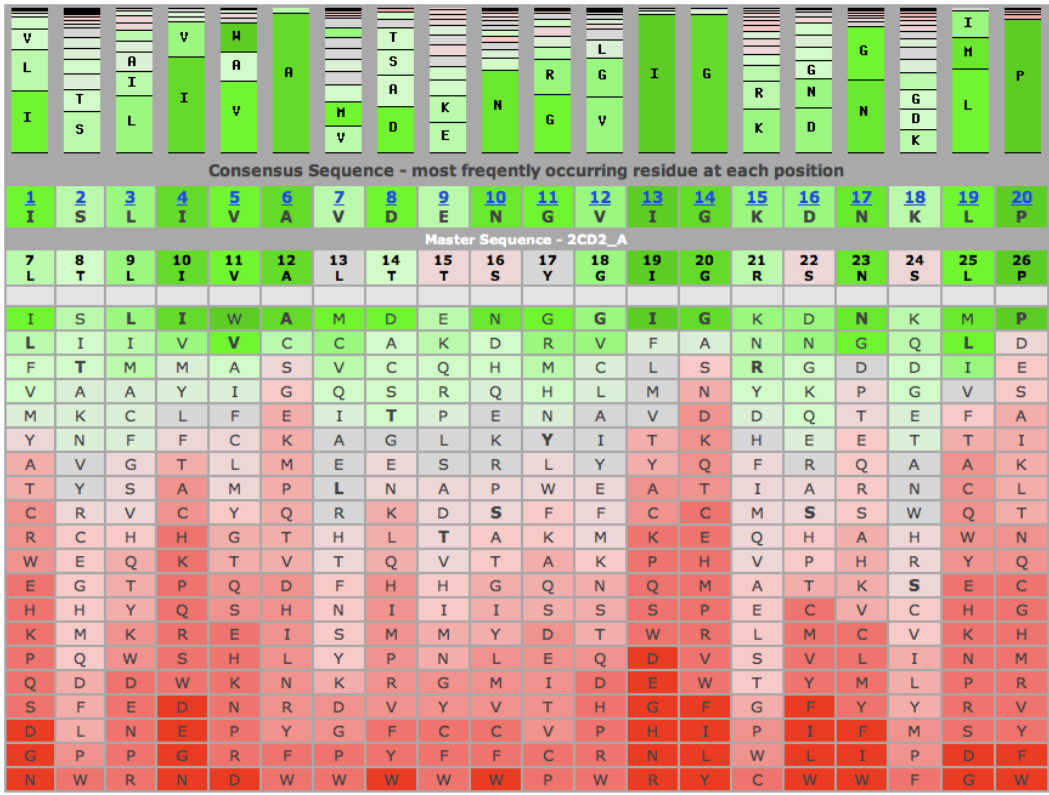
Scoring Matrix (BLOSUM62)

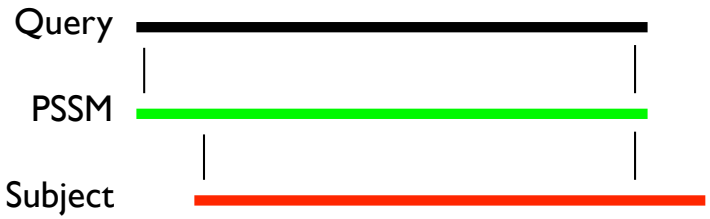
20 × 20 table

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

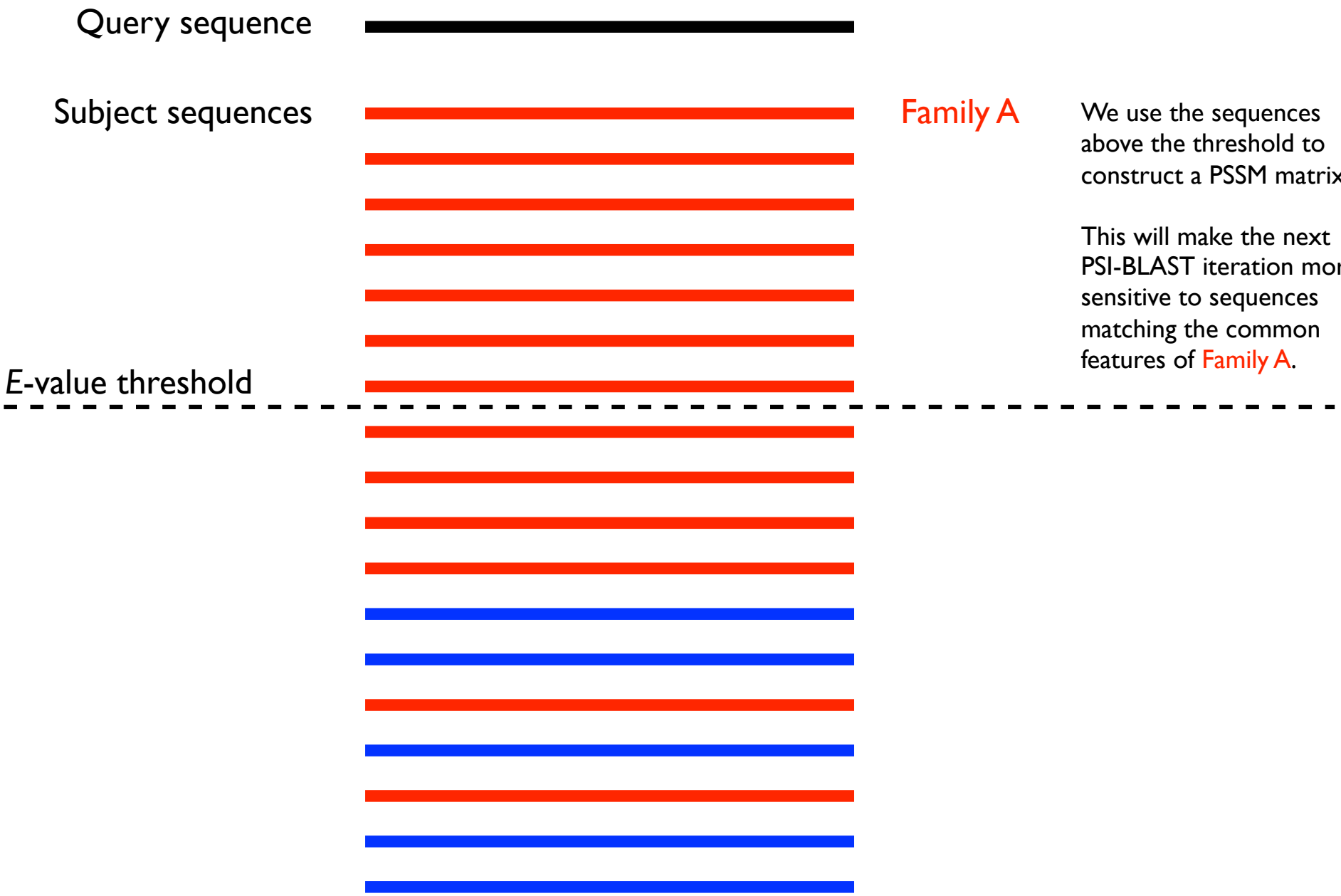
Position-Specific Scoring Matrix (PSSM)

20 × L table

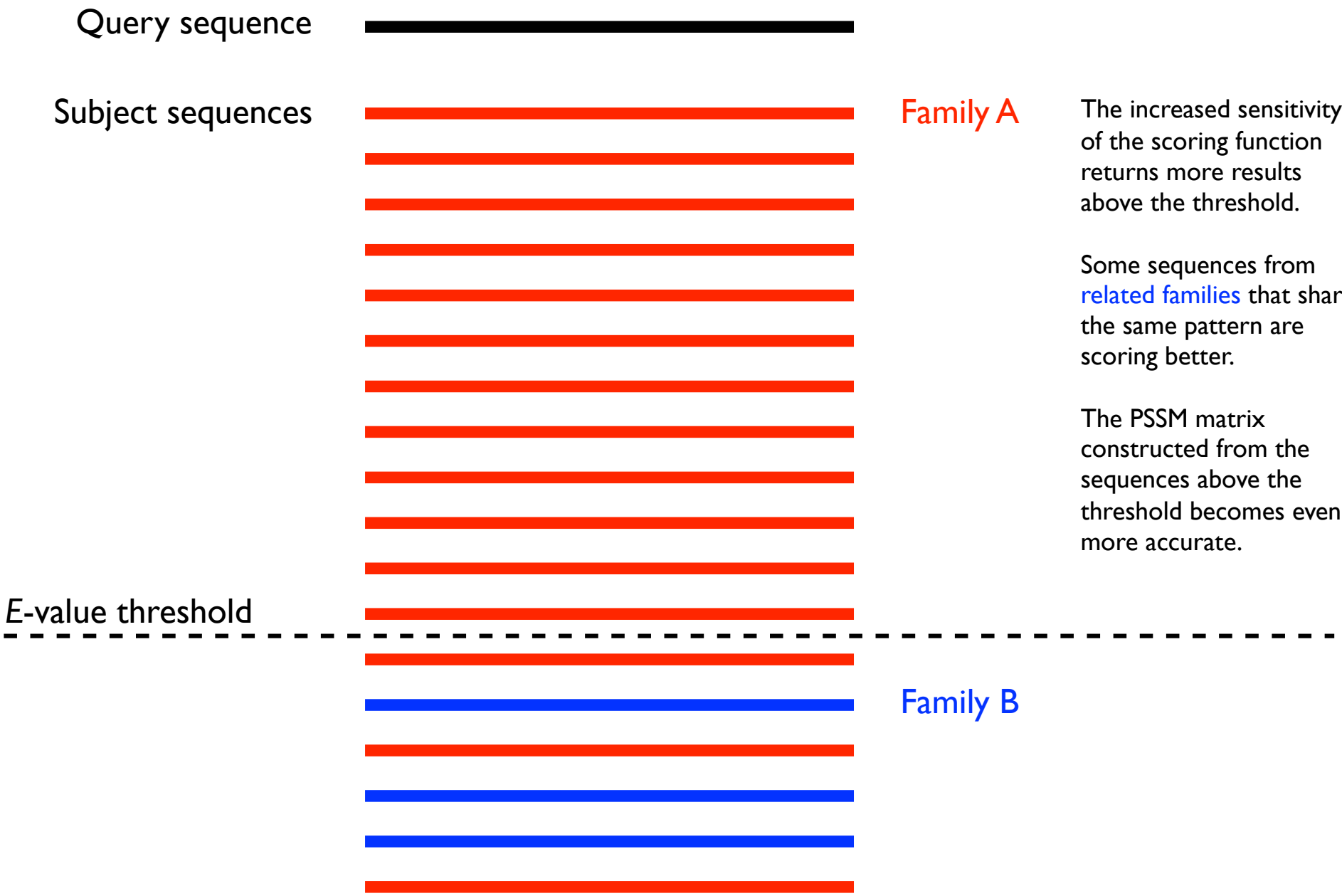




PSI-BLAST: Iteration #1



PSI-BLAST: Iteration #2

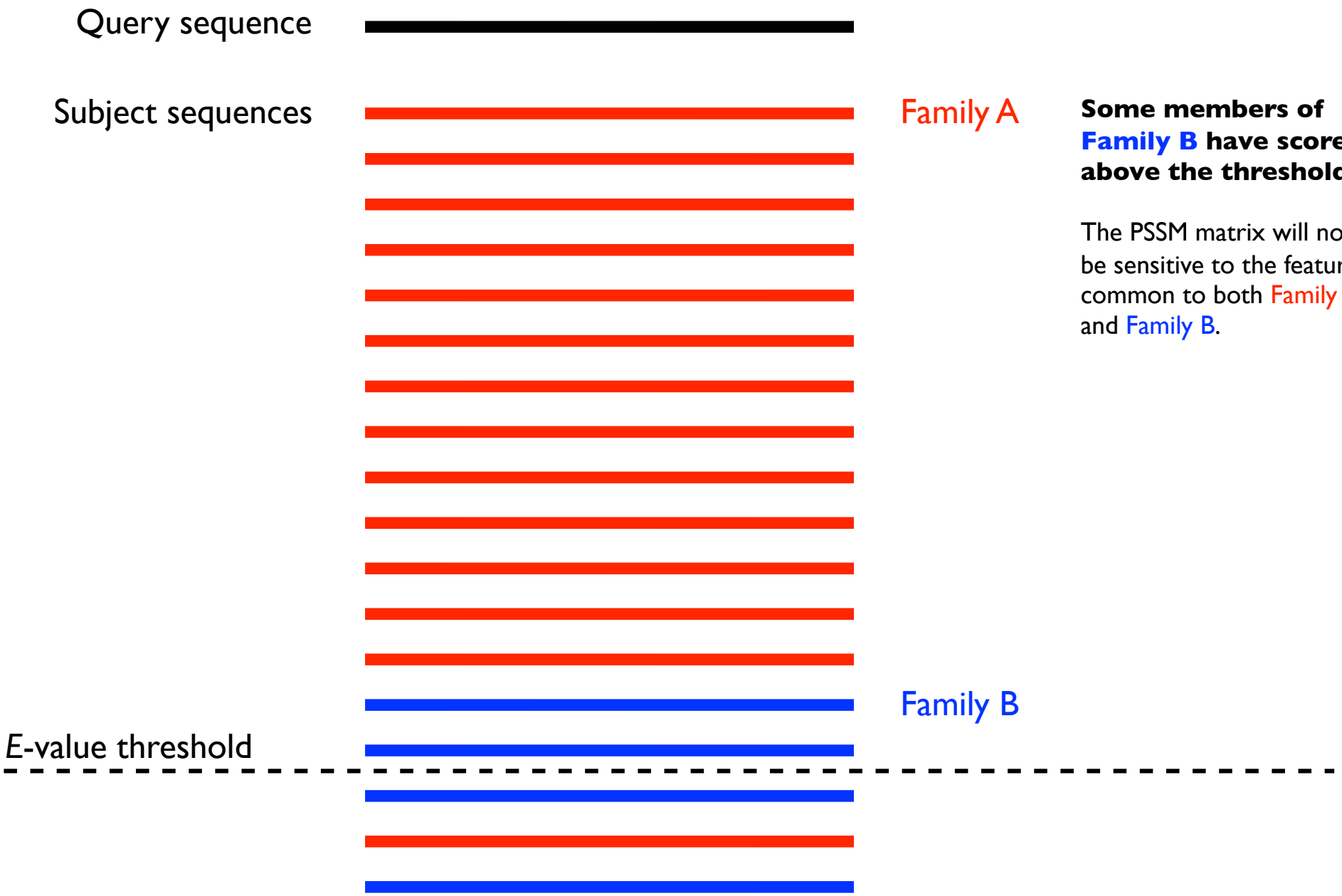


The increased sensitivity of the scoring function returns more results above the threshold.

Some sequences from related families that share the same pattern are scoring better.

The PSSM matrix constructed from the sequences above the threshold becomes even more accurate.

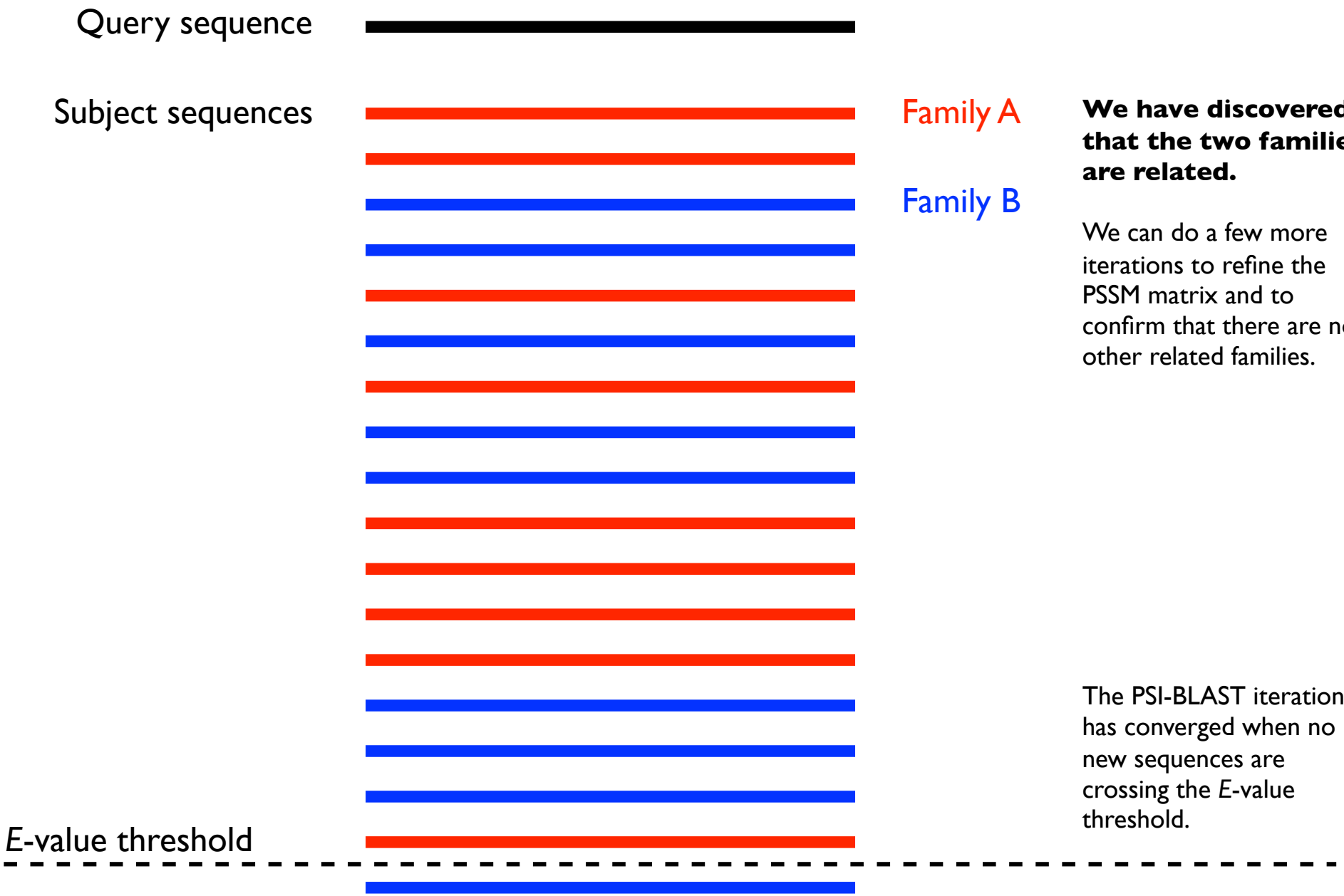
PSI-BLAST: Iteration #3



Some members of Family B have scores above the threshold.

The PSSM matrix will now be sensitive to the features common to both Family A and Family B.

PSI-BLAST: Iteration #4



We have discovered that the two families are related.

We can do a few more iterations to refine the PSSM matrix and to confirm that there are no other related families.

The PSI-BLAST iteration has converged when no new sequences are crossing the *E*-value threshold.

PSI-BLAST: "Issues"

What if there is a Family C in the picture?

Depending on which of Family B or Family C gets above the threshold first, the PSSM will become more sensitive to either the common features of A and B or of A and C. As soon as we reach that point, the "losing" family will have its score going down.

We have a "winner-takes-all" situation.

The result of the "race" depends on the *E*-value threshold.

How do we choose the *E*-value threshold?

A higher *E*-value will produce a more "inclusive" pattern, that can be used to detect weaker homologies.

A lower one will keep the PSI-BLAST search closer to the query sequence.

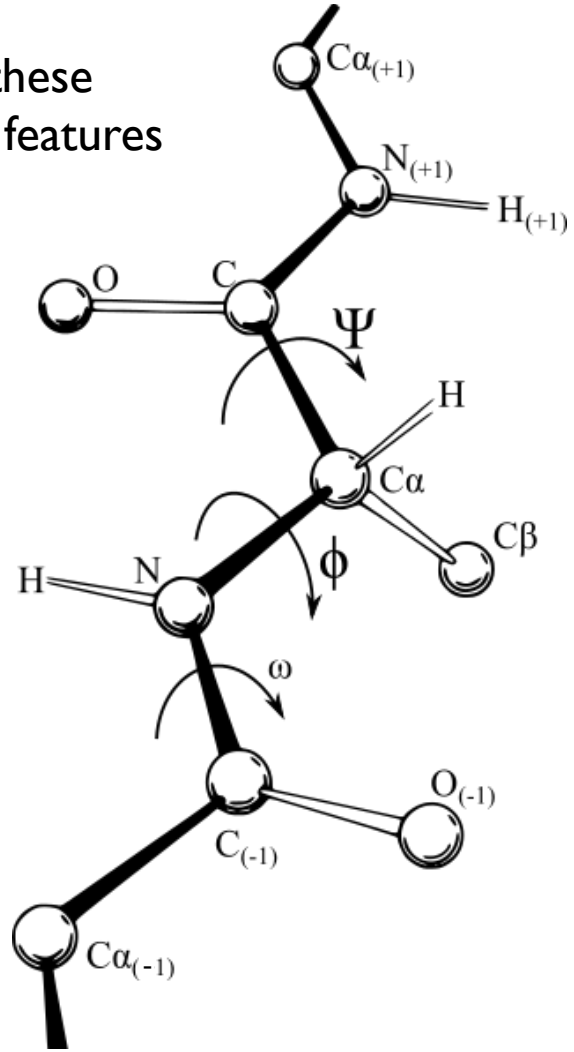
The NCBI's PSI-BLAST interface allows one to manually select sequences to be part of the PSSM construction, whether they are in the list or not.

This allows to "seed" the PSSM for certain domains we wish to discover.

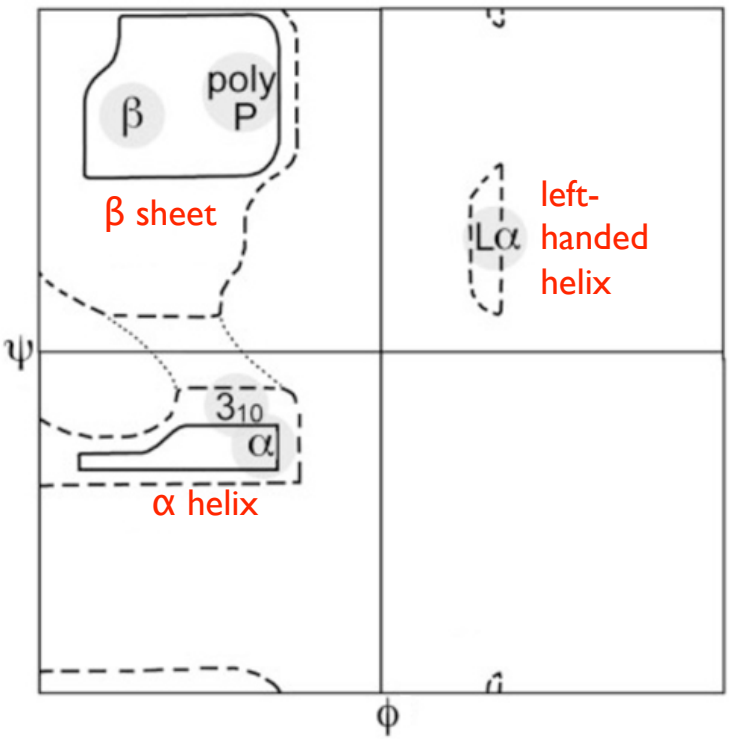
Secondary structure prediction

Is there a way to tell if a certain stretch of protein sequence is forming an **alpha-helix**, a **beta-strand**, or a **loop**?

How are these structural features defined?



Ramachandran plot (a.k.a "ϕψ plot")



Source: **Wikipedia**
http://en.wikipedia.org/wiki/Ramachandran_plot

```
  AA: KGYLNTFGLATSLFVPIVEEGIFEGAILDSTIAHYLKQYPSTPNAIILGC
Pred: CCCCEEEECCCHHHHHHHHHCCCCCHHHHHHHHHHHHHCCCCCCEEEECC
Conf: 69976995289799999971424898899999986432479989999 89
```

H = helical (“alpha-helix”)
E = extended (“beta-strand”)
C = coiled (“loop”)

We can compare the accuracy of different prediction methods by calculating how well they do on known protein structures.

Assuming we distinguish 3 possible states for the secondary structure (helical, extended, or coiled), we can calculate the **Q_3 score**, which is the fraction of residues that the method correctly predicts.

The Chou-Fasman method predicts secondary structure (H, E, or C) by considering the average propensities over a stretch of 3–6 amino acids.

Chou-Fasman table :

Name	P(a)	P(b)	P(turn)
Alanine	**1.42**	0.83	0.66
Arginine	0.98	0.93	0.95
Aspartic Acid	1.01	0.54	**1.46**
Asparagine	0.67	0.89	**1.56**
Cysteine	0.70	1.19	1.19
Glutamic Acid	**1.51**	0.37	0.74
Glutamine	1.11	1.10	0.98
Glycine	**0.57**	0.75	**1.56**
Histidine	1.00	0.87	0.95
Isoleucine	1.08	**1.60**	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	**1.45**	1.05	0.60
Phenylalanine	1.13	**1.38**	0.60
Proline	**0.57**	0.55	**1.52**
Serine	0.77	0.75	**1.43**
Threonine	0.83	1.19	0.96
Tryptophan	1.08	**1.37**	0.96
Tyrosine	0.69	**1.47**	1.14
Valine	1.06	**1.70**	0.50

The P(a) values ("alpha-helix propensities") reflect the fact that Gly and Pro are considered "helix breakers".)

"a" = alpha-helix
"b" = beta-strand
"turn" = beta-turn

P > 1.00 means the secondary structure happens more often than expected.

P < 1.00 means the secondary structure happens less often than expected.

See details of the algorithm at
http://swift.cmbi.ru.nl/teach/aainfo/chou.shtml

Methods based on the propensity of individual AAs to form a certain secondary structure are usually not doing better than Q_3 = 55%.

Why is it not doing better?

Secondary structure is a collective property.

Examples:

Mezei. Chameleon sequences in the PDB. *Protein Eng.* 1998, **11**, 411–414. http://dx.doi.org/10.1093/protein/11.6.411

Minor & Kim. Context-dependent secondary structure formation of a designed protein sequence. *Nature* 1996, **380**, 730–734. http://dx.doi.org/10.1038/380730a0

To improve the performance, we have to focus on the evolutionary conserved segments.

The **PSIPRED** method finds those evolutionary conserved segments using PSI-BLAST.

Reference:
http://dx.doi.org/10.1006/jmbi.1999.3091

For PSIPRED 3.2 : Q_3 = 82%

What is still missing?

Secondary structure is not just about the local sequence. It is affected by the way the protein folds as a whole.

Presuming it folds in a particular way, of course… Some proteins are "intrinsically disordered".

Plus, who says there is just one secondary structure?