CHEM 498Q / 630Q

Molecular modelling of proteins

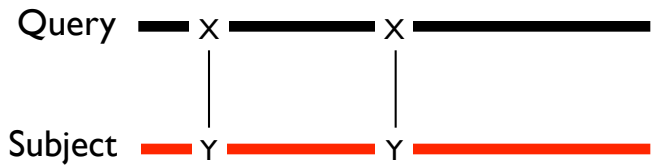
Fall 2015 Term

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.

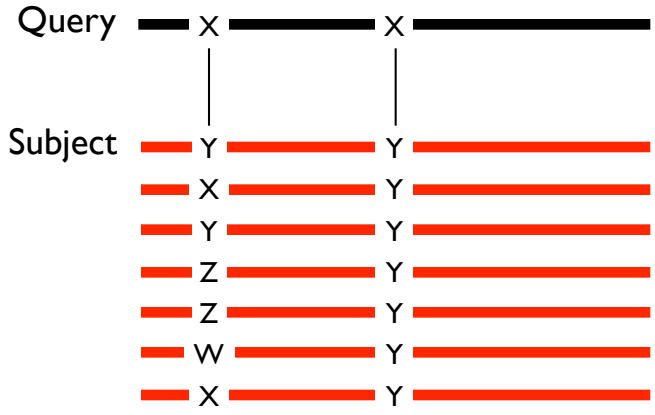
This is arguably the best we can do if we don't know what are the important domains of the query sequence.



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.



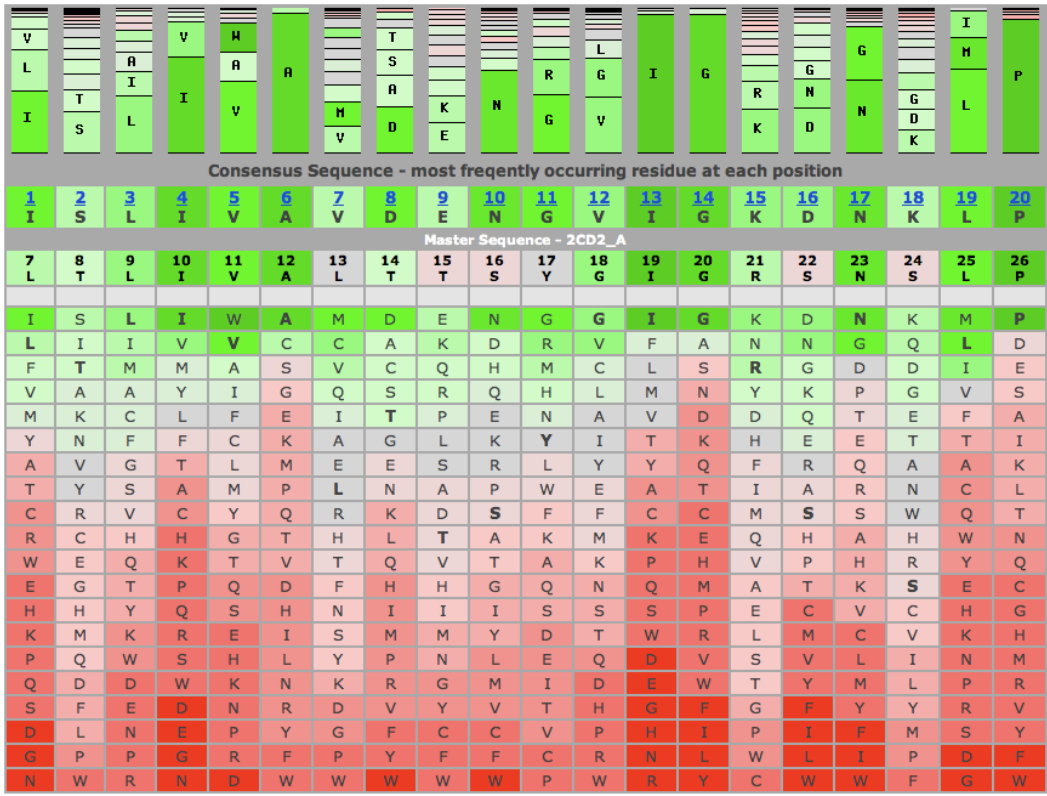
Scoring Matrix (BLOSUM62)

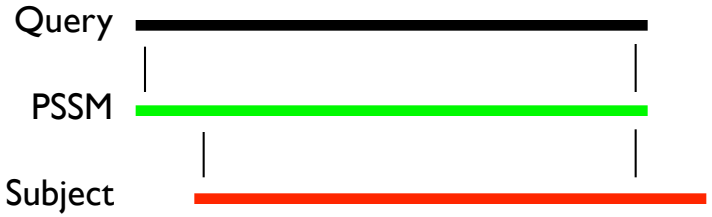
20 × 20 table

	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

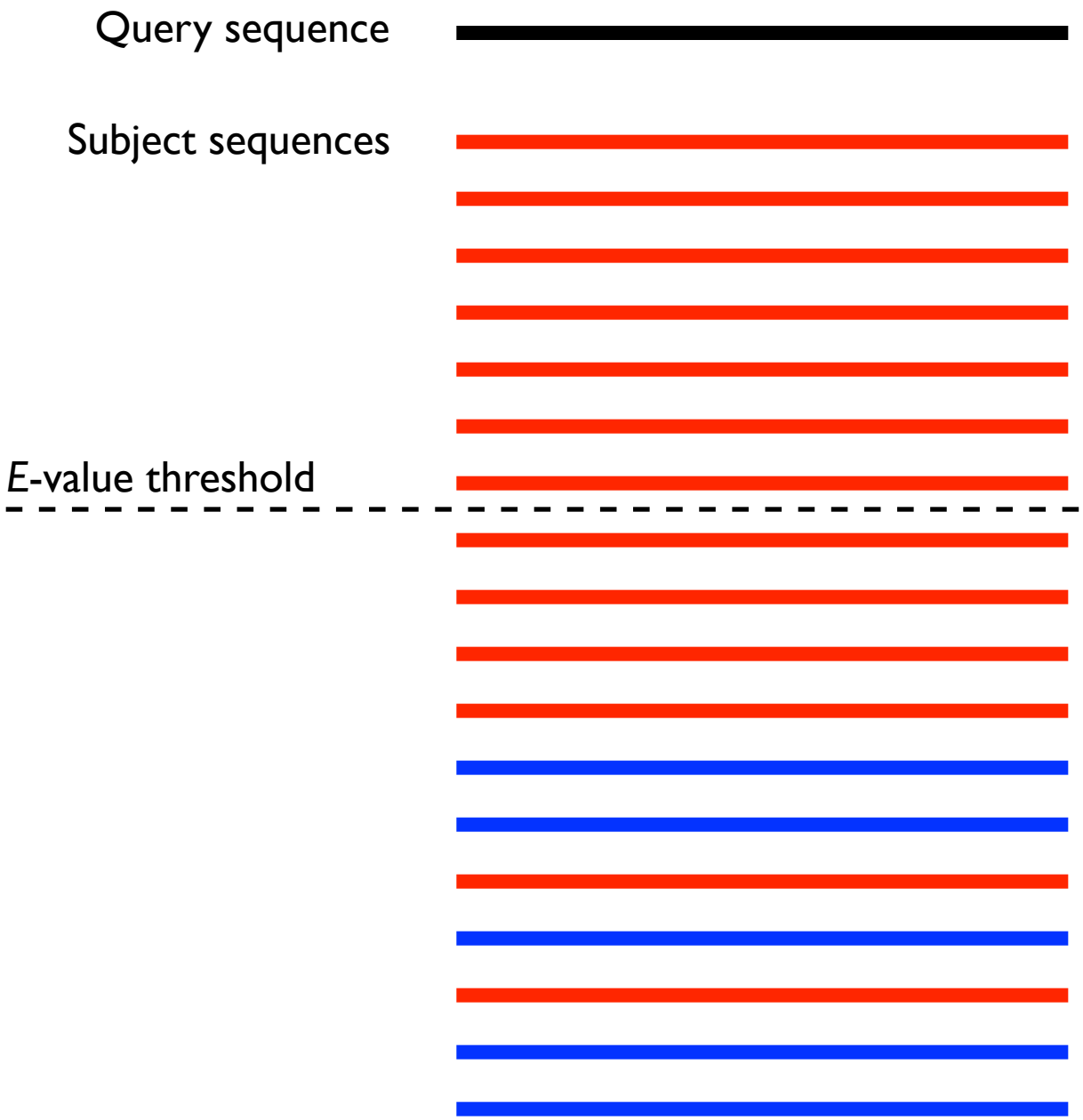
Position-Specific Scoring Matrix (PSSM)

20 × L table





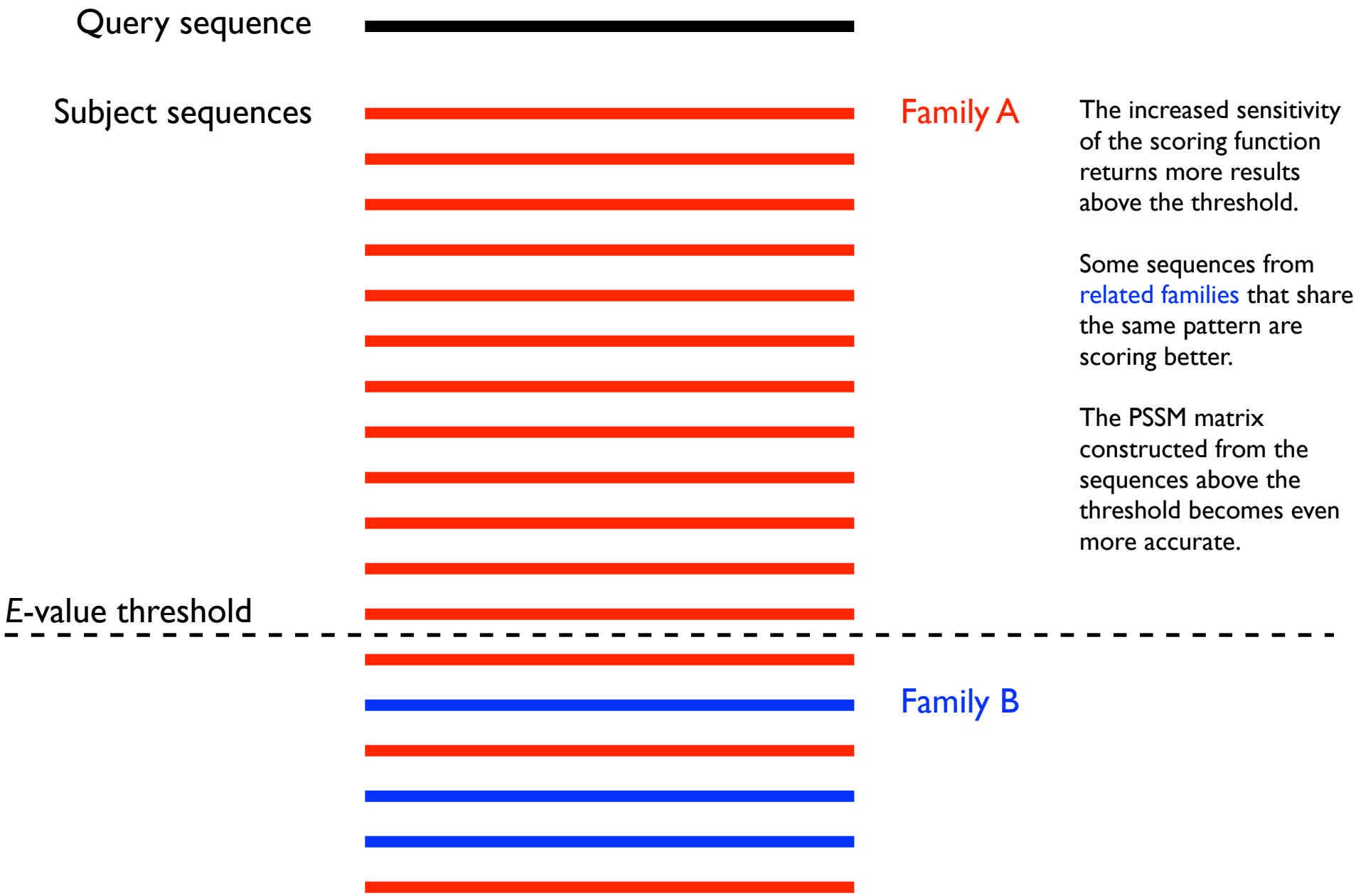
PSI-BLAST: Iteration #1



We use the sequences above the threshold to construct a PSSM matrix.

This will make the next PSI-BLAST iteration more sensitive to sequences matching the common features of Family A.

PSI-BLAST: Iteration #2

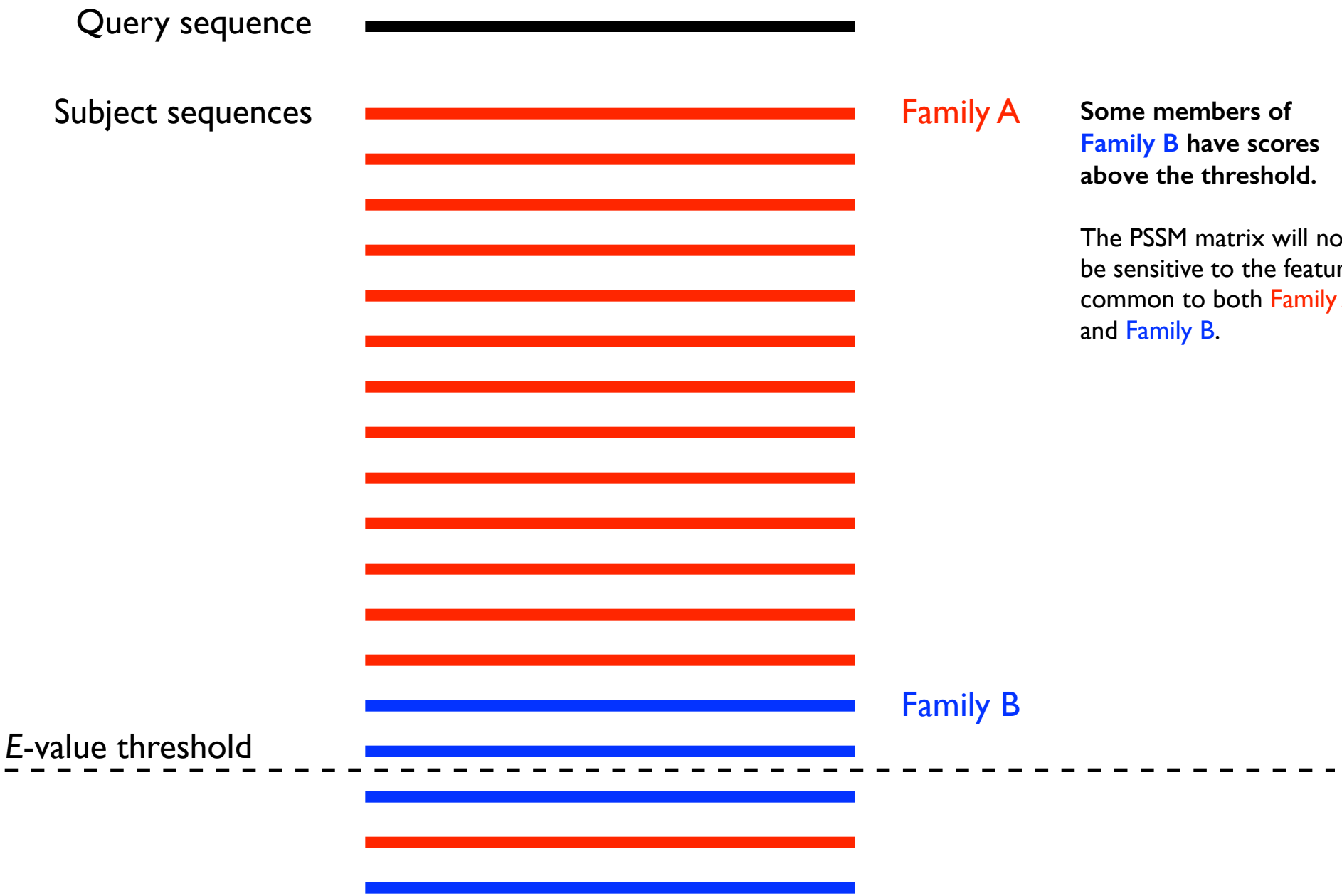


The increased sensitivity of the scoring function returns more results above the threshold.

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.

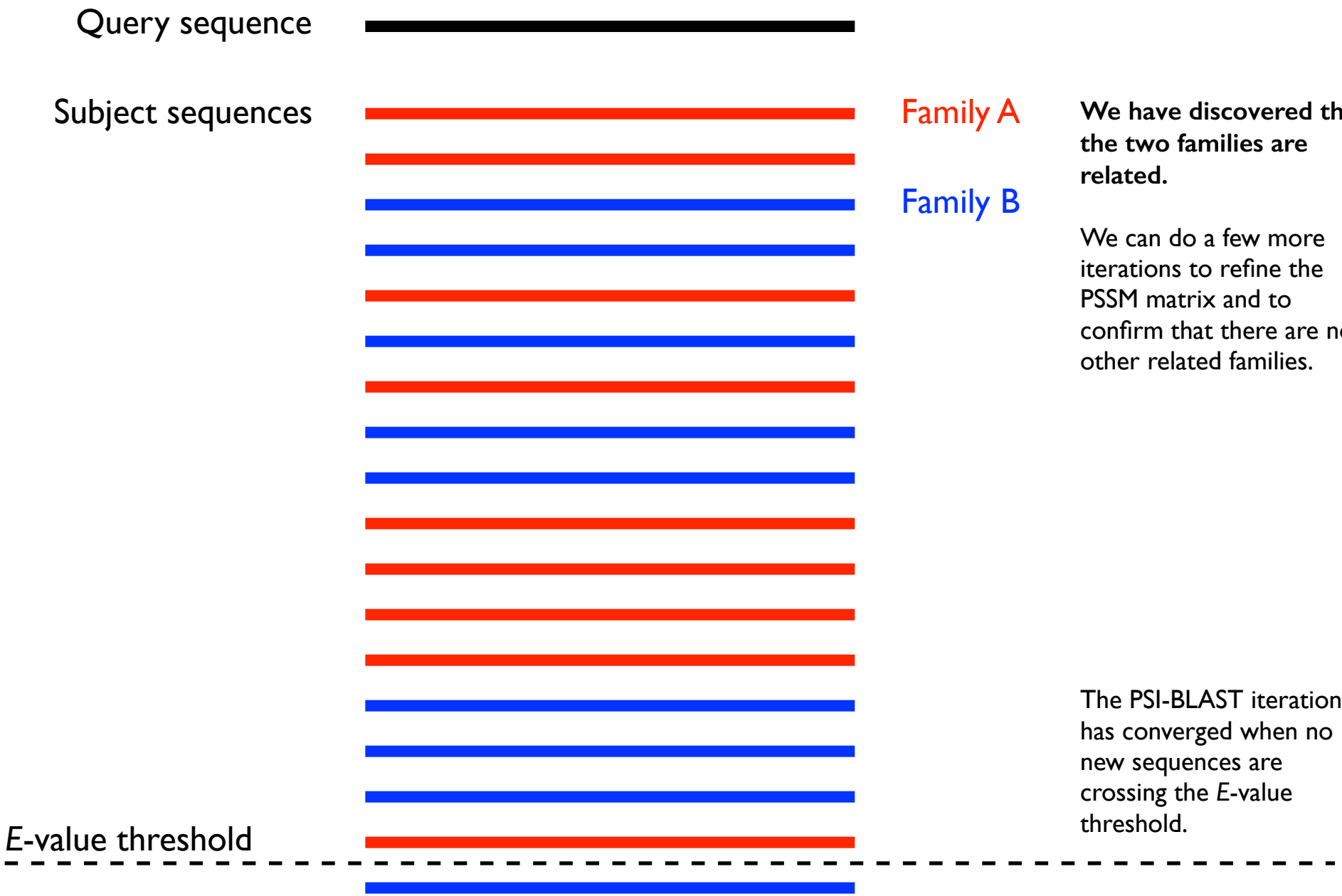
PSI-BLAST: Iteration #3



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.

Interlude…

“How George Dantzig solved the diet problem”

IRV (Irvin Lustig)
What were the first problems that were solved by the simplex method by hand or on a computer?

GEORGE (George Dantzig)
I have a good description in my book "Linear Programming and Extensions". It says: "One of the first applications of the simplex algorithm was to the determination of an adequate diet that was of least cost. In the fall of 1947, Jack Laderman of the Mathematical Tables Project of the National Bureau of Standards undertook, as a test of the newly proposed simplex method, the first large-scale computation in this field. It was a system with nine equations in seventy-seven unknowns. Using hand-operated desk calculators, approximately 120 man-days were required to obtain a solution." "The particular problem solved was one which had been studied earlier by George Stigler (who later became a Nobel Laureate) who proposed a solution based on the substitution of certain foods by others which gave more nutrition per dollar. He then examined a "handful" of the possible 510 ways to combine the selected foods. He did not claim the solution to be the cheapest but gave his reasons for believing that the cost per annum could not be reduced by more than a few dollars. Indeed, it turned out that Stigler's solution (expressed in 1945 dollars) was only 24 cents higher than the true minimum per year $39.69."

Source: **George Dantzig Memorial Site**
http://www2.informs.org/History/dantzig/in_interview_irv8.htm

George Dantzig, father of the simplex method

Interlude…

"How George Dantzig solved the diet problem"

IRV
When you solved this problem, was pivoting done by hand?

GEORGE
Yes. All computations were done using hand calculators.

IRV
I imagine there must have been a large team of people doing the arithmetic?

GEORGE
Yes. Perhaps a team of 10 people.

My Hungarian friend Andrew Vazsonyi wrote a sketch about how my wife upstaged me.

IRV
I think I heard the story, but would you mind retelling it?

George Dantzig, father of the simplex method

Source: **George Dantzig Memorial Site**
http://www2.informs.org/History/dantzig/in_interview_irv8.htm

Interlude…

“How George Dantzig solved the diet problem”

GEORGE
This is a cartoon drawn by my friend Andrew. If you look at this picture here it's pretty crude. You see vinegar, molasses, bran, and bouillon, meaning bouillon cubes, and you see here the simplex method and you see my wife's name Ann Dantzig spelled wrong. She spells it with an -e like the French do. The story is that I left the Pentagon in 1952 and took a position with the Rand Corporation. I decided to use the simplex method and linear programming to solve a diet problem designed for me to lose weight. Anne agreed she would prepare my meals according to what the computer declared was the optimal diet. So I called up my wife one day and said: "We've solved the diet problem for me, George Dantzig, and I want you to be ready when we run it, to cook supper and prepare it according to whatever the computer says."

So it's getting late in the day and finally Anne calls me up and she says: "Well, what's for supper?" And I said: "Well, we ran the program. A couple of gallons of vinegar and some other stuff were the optimum diet. We'll just gonna have to take vinegar now as our food." Of course, we decided vinegar was not a food. The next day, we deleted vinegar from the list of foods eligible to be in the diet and it found an optimal diet containing, among other foods, 200 bouillon cubes.

Source: **George Dantzig Memorial Site**
http://www2.informs.org/History/dantzig/in_interview_irv8.htm

George Dantzig, father of the simplex method

Interlude…

“How George Dantzig solved the diet problem”

IRV
Cubes dissolved in a cup of hot water to make a cup of soup?

GEORGE
Yes. Anne said: "Well, I'll buy the very best bouillon cubes money can buy, but be prepared to go to the hospital. I decided to start with five per day and gradually work up to a couple hundred bouillon cubes per day. Have you ever tasted five bouillon cubes dissolved in a cup of hot water?

IRV
No. What does it taste like?

GEORGE
It tastes like pure brine. I decided that two bouillon cubes was a proper upper limit per day. Each day we either eliminated or placed an upper bound on the amount of some other food in the diet.

IRV
You solved a new optimization problem each day by hand!

GEORGE
Not by hand. We were solving them on a RAND computer.

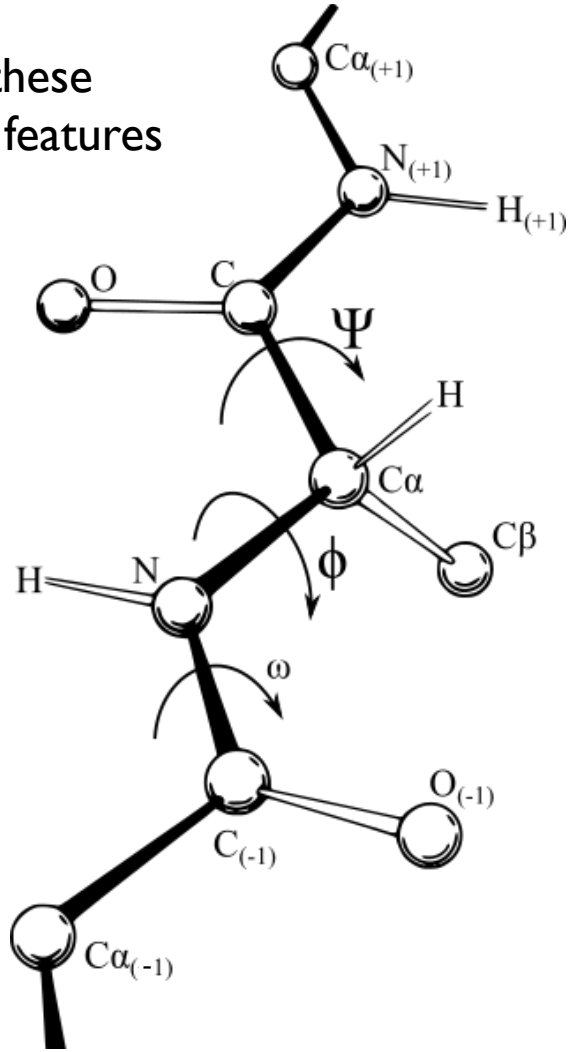
Source: **George Dantzig Memorial Site**
http://www2.informs.org/History/dantzig/in_interview_irv8.htm

George Dantzig, father of the simplex method

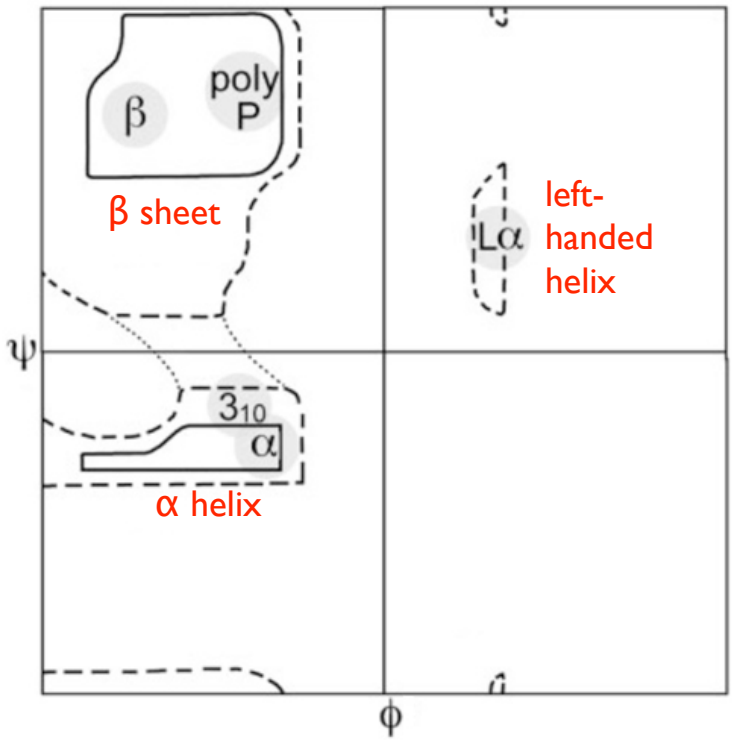
Secondary structure prediction

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: 699769952897999999714248988999999864324799899989
```

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.

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

What Q_3 score can we expect from chance alone?

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

More successful methods can be devised by considering the average propensities over a stretch of 3–6 amino acids.

Chou-Fasman table :

Name	P(a)	P(b)	P(turn)
Alanine	**142**	83	66
Arginine	98	93	95
Aspartic Acid	101	54	**146**
Asparagine	67	89	156
Cysteine	70	119	119
Glutamic Acid	**151**	037	74
Glutamine	111	110	98
Glycine	57	75	156
Histidine	100	87	95
Isoleucine	108	**160**	47
Leucine	121	130	59
Lysine	**114**	74	101
Methionine	**145**	105	60
Phenylalanine	113	**138**	60
Proline	57	55	152
Serine	77	75	143
Threonine	83	**119**	96
Tryptophan	108	**137**	96
Tyrosine	69	**147**	114
Valine	106	**170**	50

See details of the algorithm at
http://prowl.rockefeller.edu/aainfo/chou.htm

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

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

For PSIPRED : $Q_3 = 72 \pm 10\%$

What is (still) the problem?

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".)