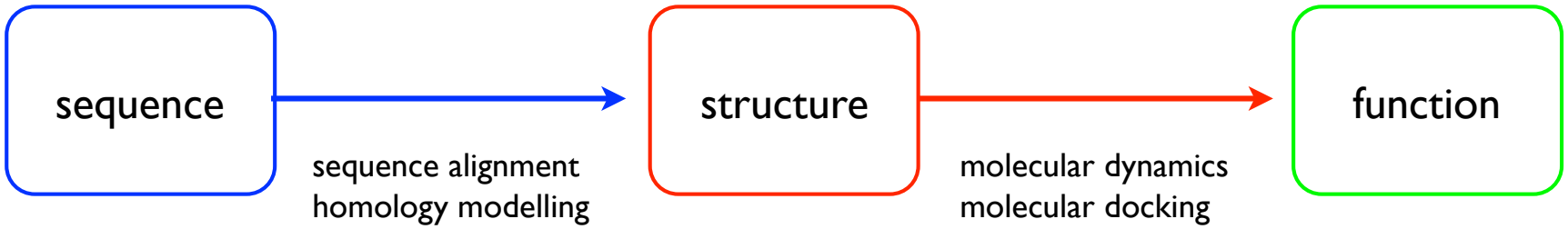
CHEM 436 / 630

Molecular modelling of proteins

Winter 2018 Term

Instructor:
Guillaume Lamoureux
Concordia University, Montréal, Canada

Overview of the course



Sources of information	**UniProt, PDB, etc.**	**CHARMM ff, Molecular libraries, etc.**
Algorithms	**BLAST, MODELLER, etc.**	**NAMD, AutoDock, etc.**

Comparing protein sequences : Similarity *versus* Homology

Identity
Proteins have the same amino acid sequence.

→ Yes or no.
(Easy to check.)

Similarity
Protein sequences have "similar" amino acids.

Nonpolar : Gly (G), Ala (A), Val (V), Leu (L), etc.
Polar : Asn (N), Ser (S), etc.
Acidic : Asp (D), Glu (E)
Basic : His (H), Lys (K), Arg (R)
etc.

→ Something we can quantify.

(We have to decide on a similarity measure, though, which has a certain degree of arbitrariness.)

Homology
Proteins are related to a common ancestor.

→ Either true or false.
(We may or may not know.)

Comparing protein sequences : Similarity *versus* Homology

Homologous proteins very often have similar structures.

Rost, *Protein Eng.* 1999, **12**, 85-94.
http://dx.doi.org/10.1093/protein/12.2.85

sequence identity > 30% means prob(homology) > 90%
sequence identity < 25% means prob(homology) < 10%

25% < identity < 30% means that you are in the "twilight zone"…

Homologous proteins don't necessarily have the same function.

Hegyi & Gertstein, *J. Mol. Biol.* 1999, **288**, 147-164.
http://dx.doi.org/10.1006/jmbi.1999.2661

Russell, Sasieni, Sternberg, *J. Mol. Biol.* 1998, **282**, 903-918.
http://dx.doi.org/10.1006/jmbi.1998.2043

Comparing protein sequences : Similarity *versus* Homology

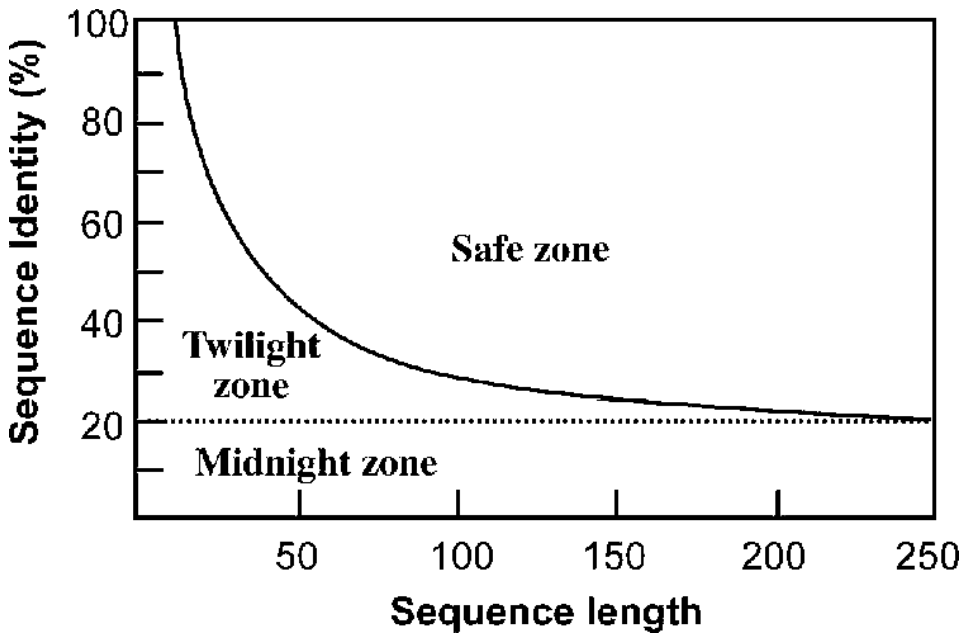


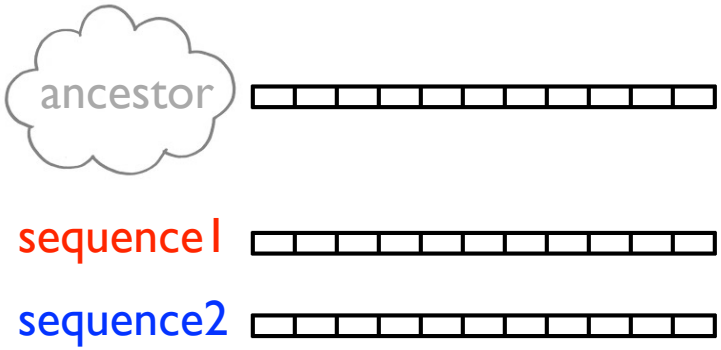
Figure 3.1: The three zones of protein sequence alignments. Two protein sequences can be regarded as homologous if the percentage sequence identity falls in the safe zone. Sequence identity values below the zone boundary, but above 20%, are considered to be in the twilight zone, where homologous relationships are less certain. The region below 20% is the midnight zone, where homologous relationships cannot be reliably determined. (*Source:* Modified from Rost 1999).

Figure from:
Jin Xong, *Essential Bioinformatics* (2006), Cambridge University Press

Sequence alignment

"Hypothesis about which pairs of residues have evolved from the same ancestral residue"

(Zvelebil & Baum, p.74)



Similarity matrix

What are the odds that a certain amino acid X in one sequence is the homolog of amino acid Y in another?

Similarity matrices S(X,Y) are usually obtained from the "log odd ratios" observed in a large number of "reference" alignments :

$$S(\mathrm{X},\mathrm{Y}) = \log_{10}\left(\frac{\#\text{ of X-to-Y transitions } / \#\text{ of transitions}}{2f(\mathrm{X})f(\mathrm{Y})}\right)$$

20×20 matrix

frequency of amino acid X

frequency of amino acid Y

The odds are high if a large number of X-to-Y transitions is observed in those alignments (relative to the total number of transitions).

PQPLEQIKISESQLAGRVGYVEMDLASGRTLAAWRASERFPLMSTFKVLLCGAVLARVDA
GDEQLDRRIHYRQQDLVDYSPVSEKHLADGMTVGELCAAAITMSDNTAGNLLLKIVGGPA
GLTAFLRQIGDNVTRLDRWETELNEALPGDVRDTTTPASMATTLRKLLTTPSLSARSQQQ
LLQWMVDDRVAGPLIRAVLPAGWFIADKTGAGERGSRGIVALLGPDGKAERIVVIYLRDT
AATMAERNQQIAGIGAALIEHWQR

PQPLEQVTRSESQLAGRVGYVEMDLASGRTLAAWRASERFPLMSTFKVLLCGAVLARVDA
GDEQLDRRIRYRQQDLVDYSPVSEKHLADGMTVGELCAAAITMSDNSAGNLLLKSVGGPA
GLTAFLRQIGDNVTRLDRWETELNEALPGDVRDTTTPASMAATLRKLLTSHALSARSQQQ
LLQWMVDDQVAGPLIRAVLPAGWFIADKTGAGERGSRGIVALLGPNGKAERIVVIYLRDT
PATMAERNQQIARIGAALIEHWQR

PQPLEQVTRSESQLAGRVGYVEMDLASGRTLAAWRASERFPLMSTFKVLLCGAVLARVDA
GDEQLDRRIRYRQQDLVDYSPVSEKHLADGMTVGELCAAAITMSDNSAGNLLLKSVGGPA
GLTAFLRQIGDNVTRLDRWETELNEALPGDVRDTTTPASMAATLRKLLTSHALSARSQQQ
LLQWMVDDQVAGPLIRAVLPAGWFIADKTGAGERGSRGIVALLGPNGKAERIVVIYLRDT
PATMAERNQQIARIGAALIEHWQR

PAM1 matrix ("point accepted mutation")

Built from alignments of very similar sequences (identity > 85%, therefore likely evolutionarily related) and normalized so that they describe the probability of having 1 point mutation per 100 amino acids.

Describes in statistical terms a certain "unit" of molecular evolution.

They can be multiplied like regular matrices. Each multiplication creates a matrix describing the effect of a longer evolution time:

PAM2 = PAM1 × PAM1

PAM3 = PAM2 × PAM1

etc.

Margaret Dayhoff

Source: http://crosstalk.cell.com/blog/more-mothers-of-science

BLOSUM matrices

Built from *ungapped* alignments of sequences (mostly from core regions of proteins, where few loops are found).

BLOSUM80
Using only alignments that have more than 80% sequence identity

The NCBI-BLAST server uses **BLOSUM62** (by default), but also **BLOSUM80** and **BLOSUM45**.

It can also use **PAM30** and **PAM70**.

In principle, similar sequences should be aligned based on "low" **PAM** matrices (small evolutionary distances) or "high" **BLOSUM** matrices (highly similar sequences).

BLOSUM62 matrix

	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

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

The (pairwise) alignment problem

To align a sequence of length *N* with a sequence of length *M* :

1. Build the *N* × *M* "scoring table" according to the chosen similarity matrix.

2. Find the top-scoring path across that table, which corresponds to the best alignment.

   This can in principle be done systematically (by trying all possible alignments and comparing their scores, using the *Needleman–Wunsch algorithm*), but most useful software rely on a *heuristic approach*.

To align two sequences of different lengths, we may have to introduce **gaps**. These gaps carry a penalty.

In BLAST using BLOSUM62, there is a default penalty of 11 units to open a gap, and of 1 unit to extend it by one amino acid.

We quantify the statistical significance of a given alignment using the ***E*-value**, which is the number of sequences one can expect to match the query sequence with a score ***S*** (or higher) due to chance alone.

In other words, the ***E*-value** is the number of alignments that may be as good or better as the one found using the heuristic method.

Clearly, there is no point in considering alignments that have *E*-values close to 1.

G	D	N	V	T	R
D	V	R	D		
	D	V	R	D	
		D	V	R	D
D	V	R		D	
	D	V	R		D
D	V		R	D	
	D	V		R	D
D		V	R	D	
	D		V	R	D
D	V	R			D
D	V			R	D
D			V	R	D

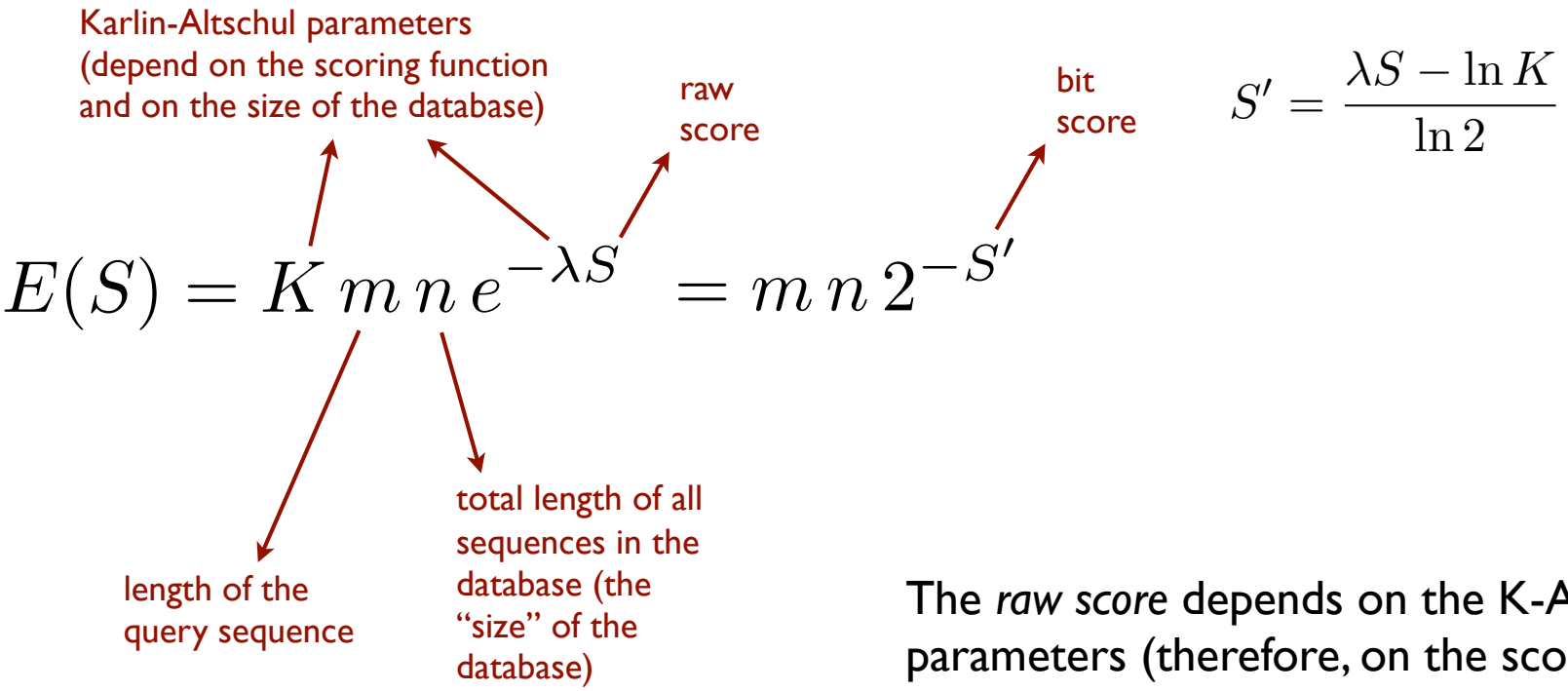
Score

Don't forget the gap penalty!

Example:
–11 for gap existence
–1 for gap extension

E-value in BLAST

The maximum score found from aligning a sequence to a *database* of sequences is presumed to follow an "extreme value distribution" :



The *raw score* depends on the K-A parameters (therefore, on the scoring function and on the size of the database) but the *bit score* does not.

For more information:
http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

To get started with BLAST

BLAST Help :
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs