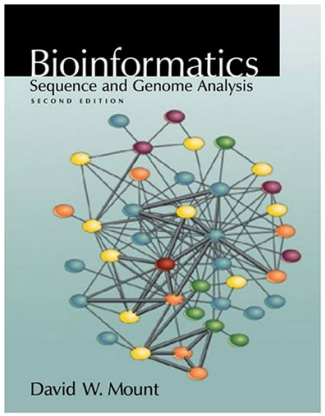
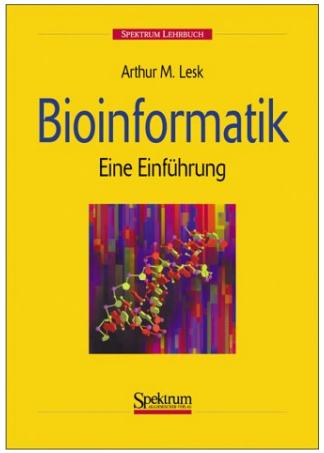
V3 - Multiple Sequence Alignment

Literature: see chapter 4 in book by David Mount



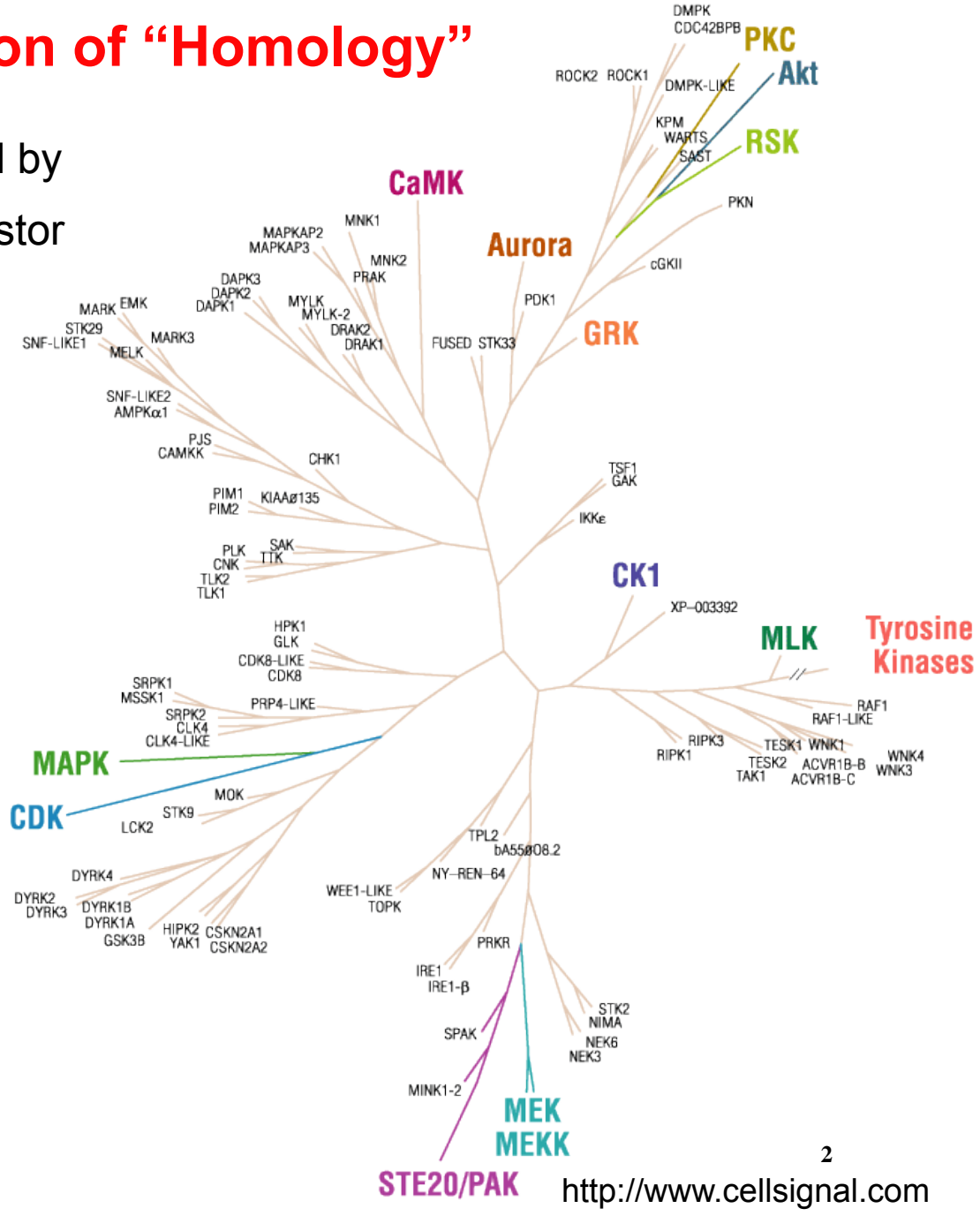
Thioredoxin-example taken from book of Arthur Lesk



Definition of “Homology”

- **Homology**: **Similarity**, caused by originating from common ancestor gene –
  identifying and analyzing homologies is the central task of the field phylogeny.

- An **Alignment** is a hypothesis for the positional homology between base pairs o amino acids.



http://www.cellsignal.com

Alignments can be simple or difficult

```
GCGGCCCA TCAGGTACTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
******** ********** *****
```

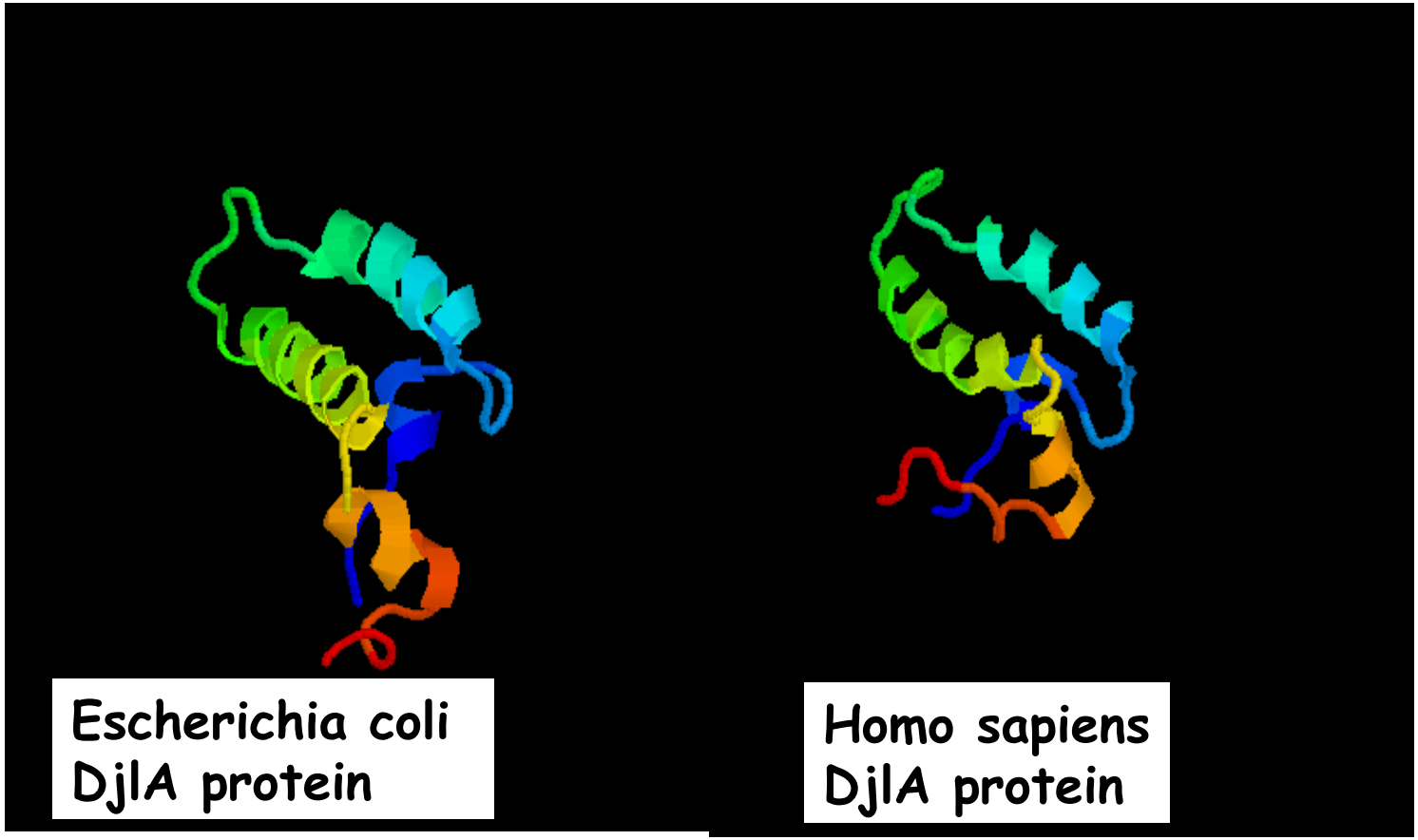
simple

```
TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
******** ?????????? *****
```

Difficult due to insertions and deletions (indels)

Can one proof whether an alignment is correct?

Protein-Alignment can be led by Information about tertiary structure

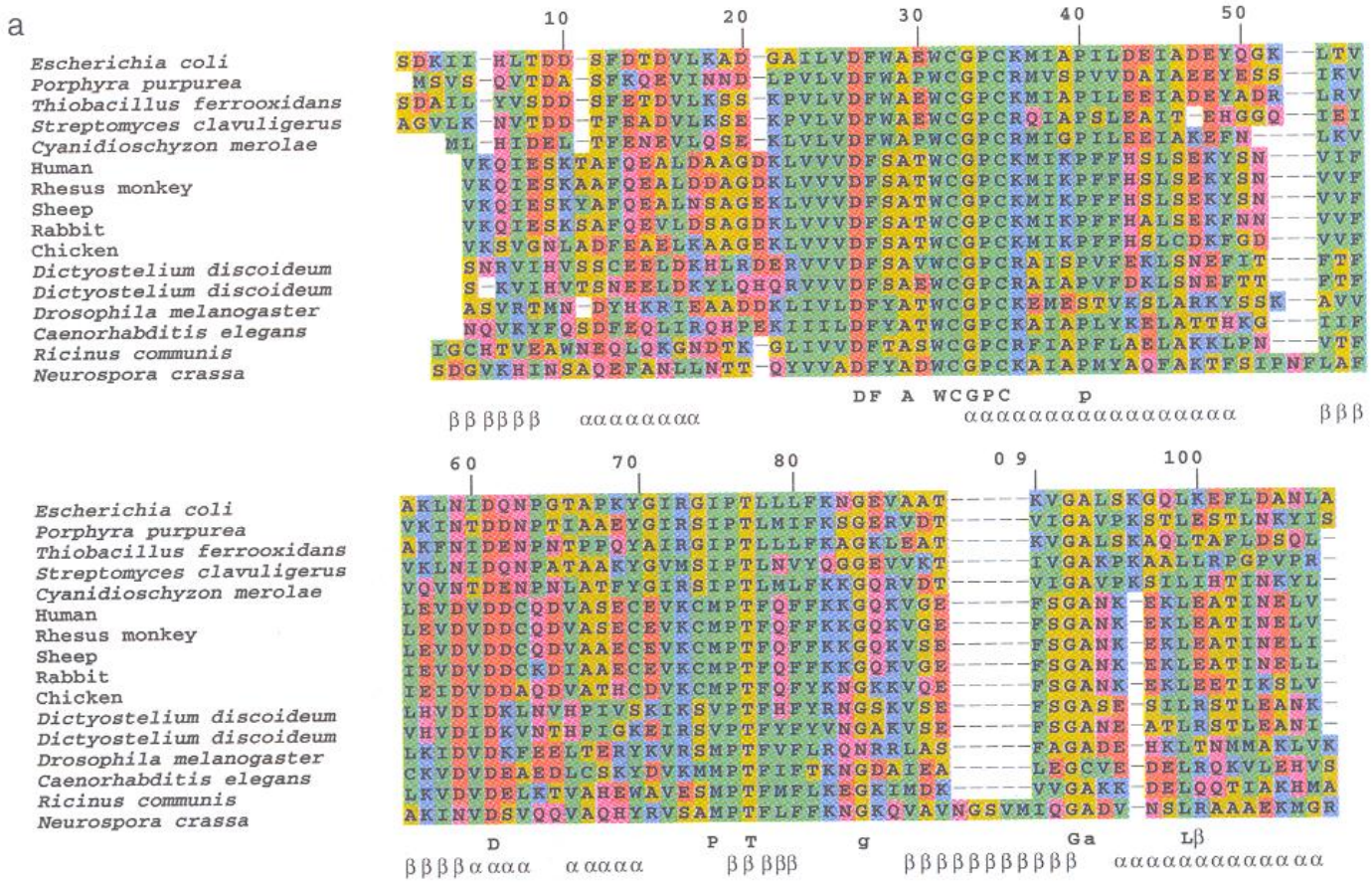


Gaps in an alignment should be preferably be placed outside secondary-structure elements

On the basis of two 3D structures one can score whether a sequence alignment is correct.
However, this is no proof in a mathematical sense.

MSA for Thioredoxin family

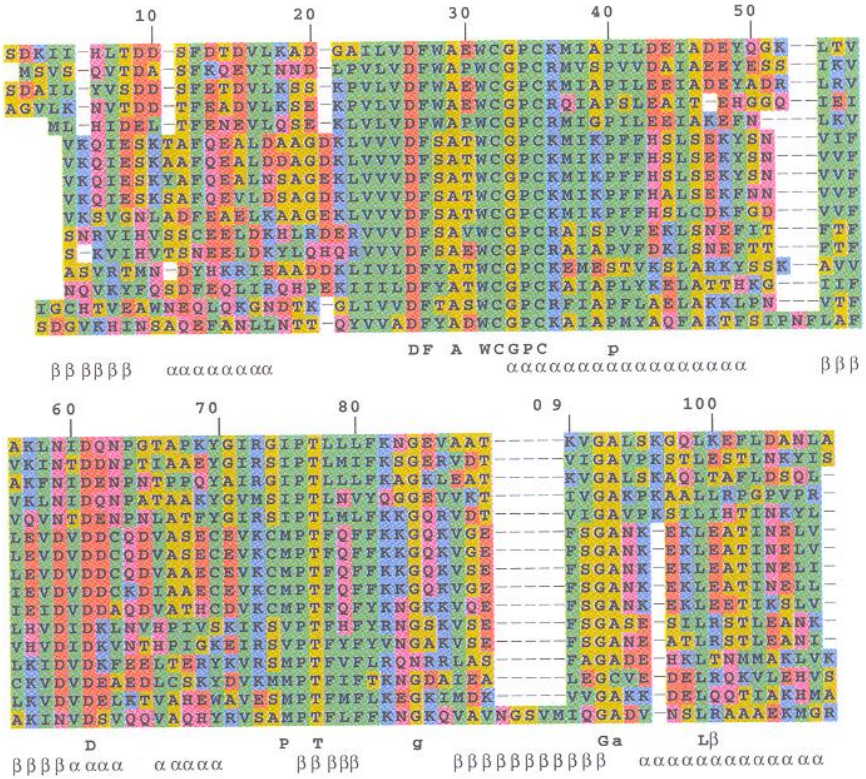
color	amino acid type	Amino acids
yellow	small, less polar	Gly, Ala, Ser, Thr
green	hydrophobic	Cys, Val, Ile, Leu
		Pro, Phe, Tyr, Met, Trp
purple	polar	Asn, Gln, His
red	negatively charged	Asp, Glu
blue	positively charged	Lys, Arg

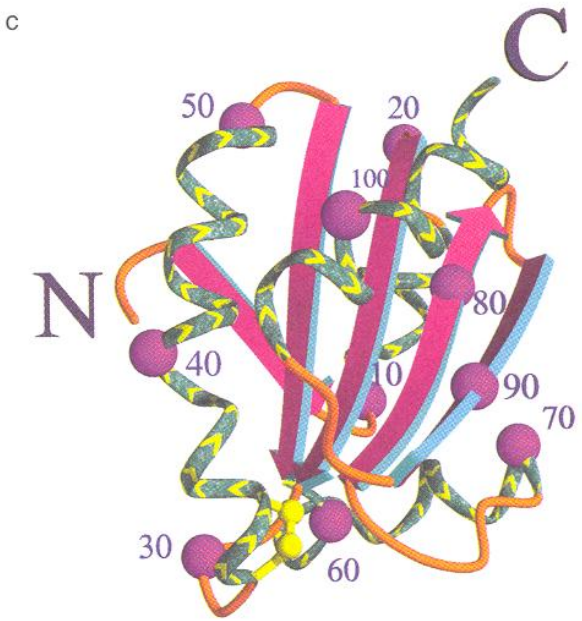


Infos from MSA of Thioredoxin family

Thioredoxin: beta-sheet formed by 5 beta strands,
flanked on both sides by alpha-helices.

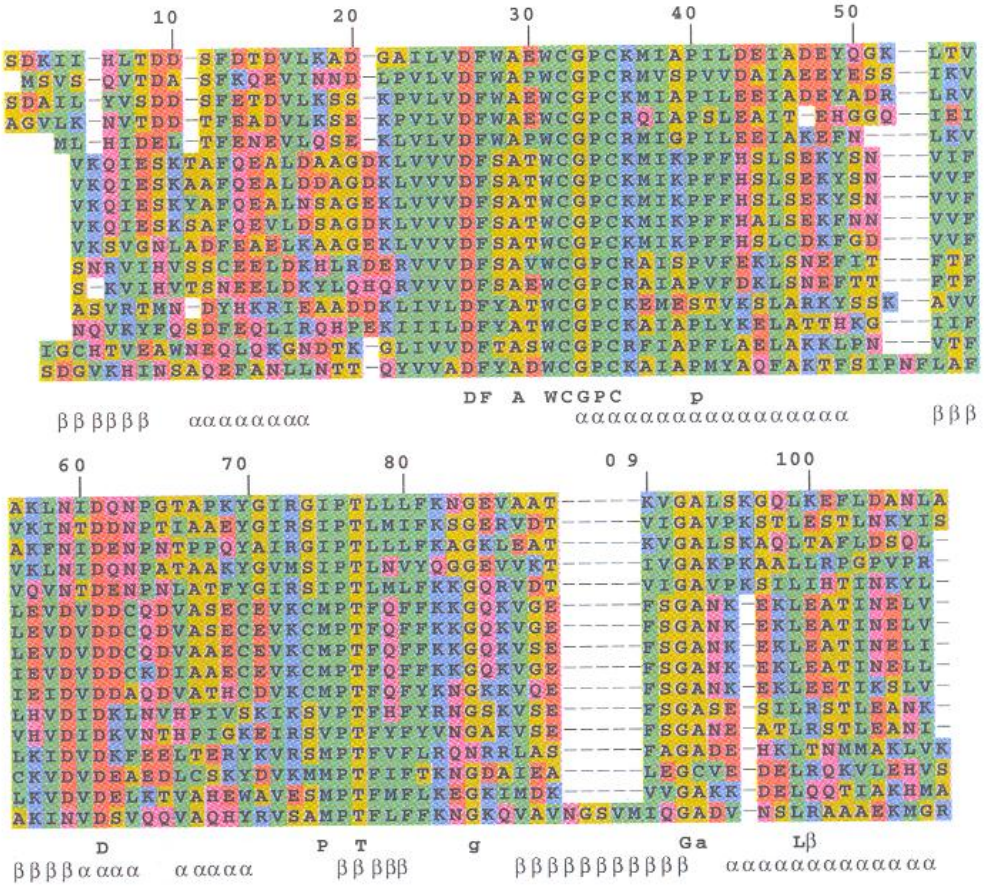
Common function: reduce disulfide bridges in other proteins

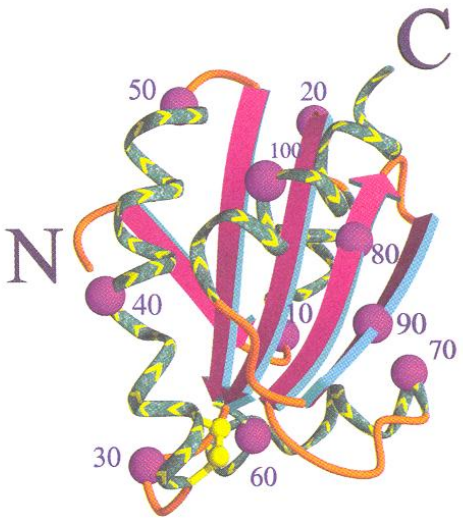




Infos from MSA of Thioredoxin family

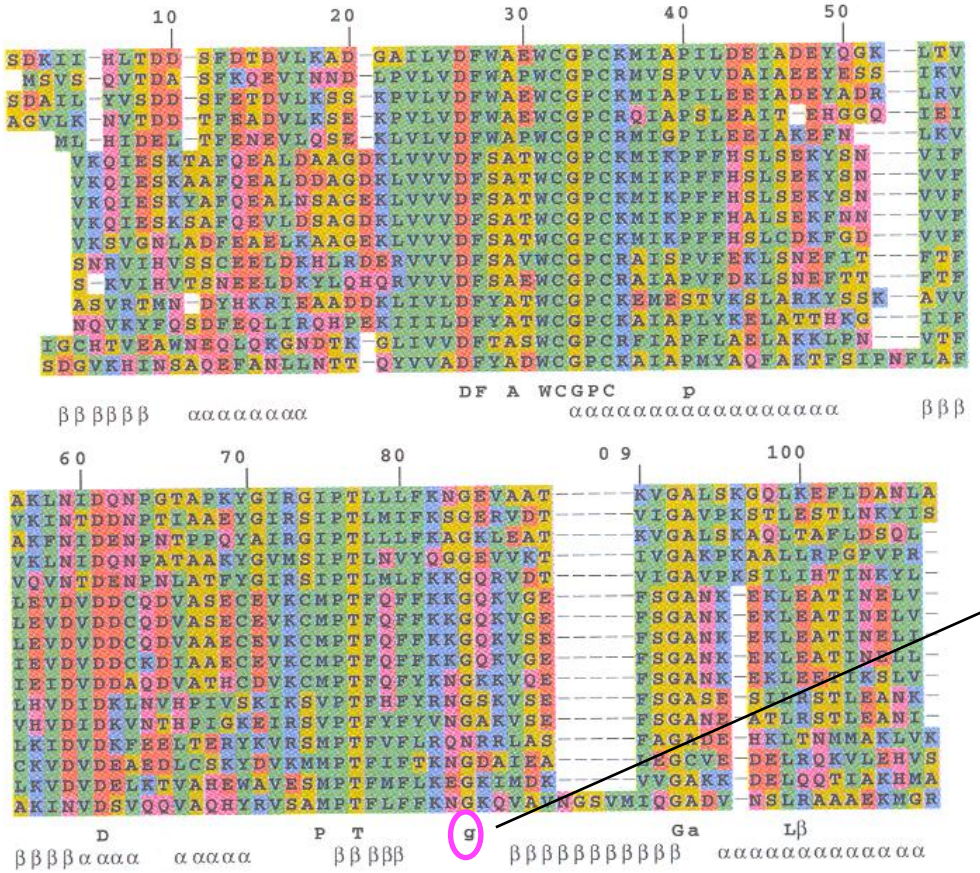
1) The most conserved parts belong typically to the active center of the protein. Here, the disulfide bridge between Cys32 und Cys35 belongs to the conserved WCGPC[K or R] motif. Other conserved stretches, e.g. Pro76Thr77 and Gly92Gly93 participate in binding the substrate.

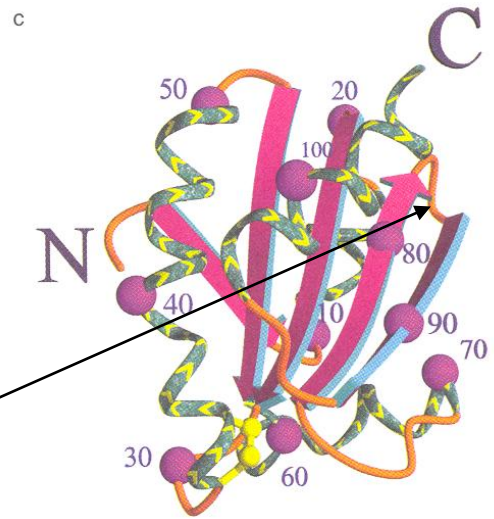




Infos from MSA of Thioredoxin family

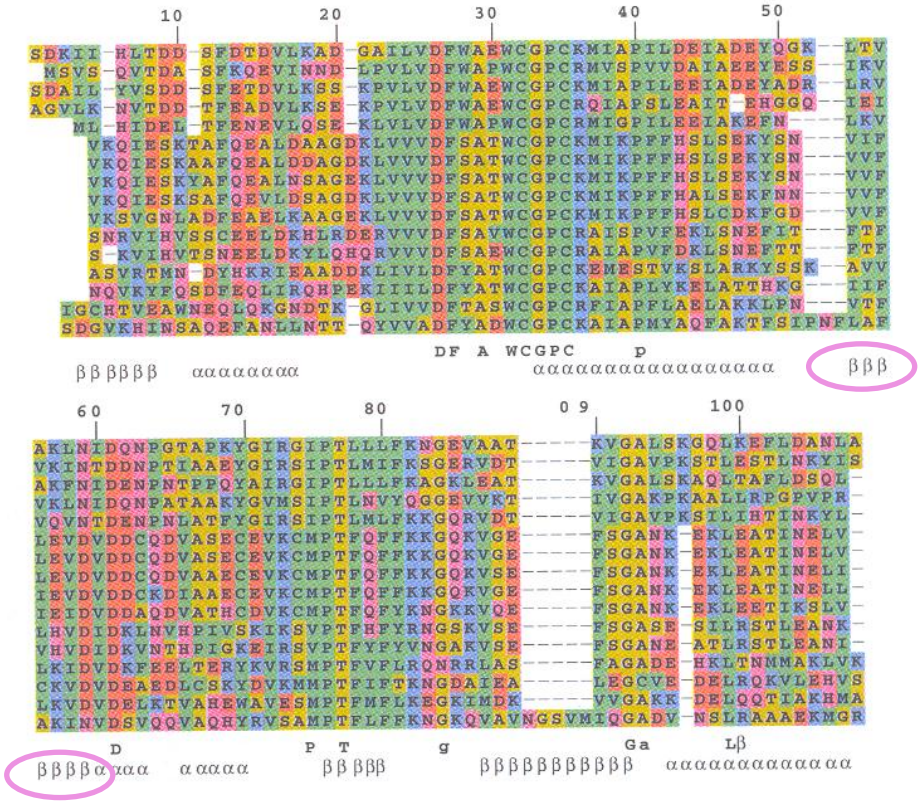
2) Stretches with many insertions and deletions likely correspond to loops on the surface. A position with a conserved Gly or Pro suggests a turn of the chain.

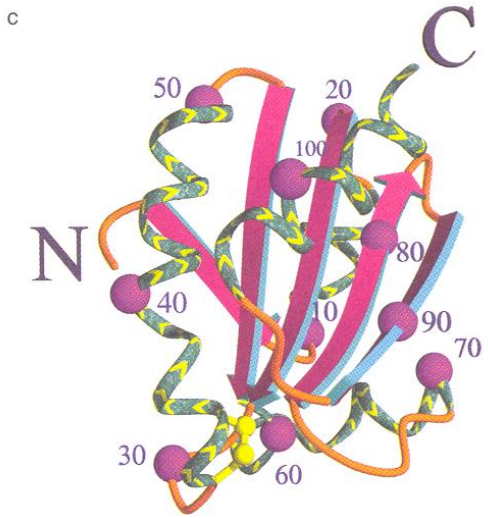




Infos from MSA of Thioredoxin family

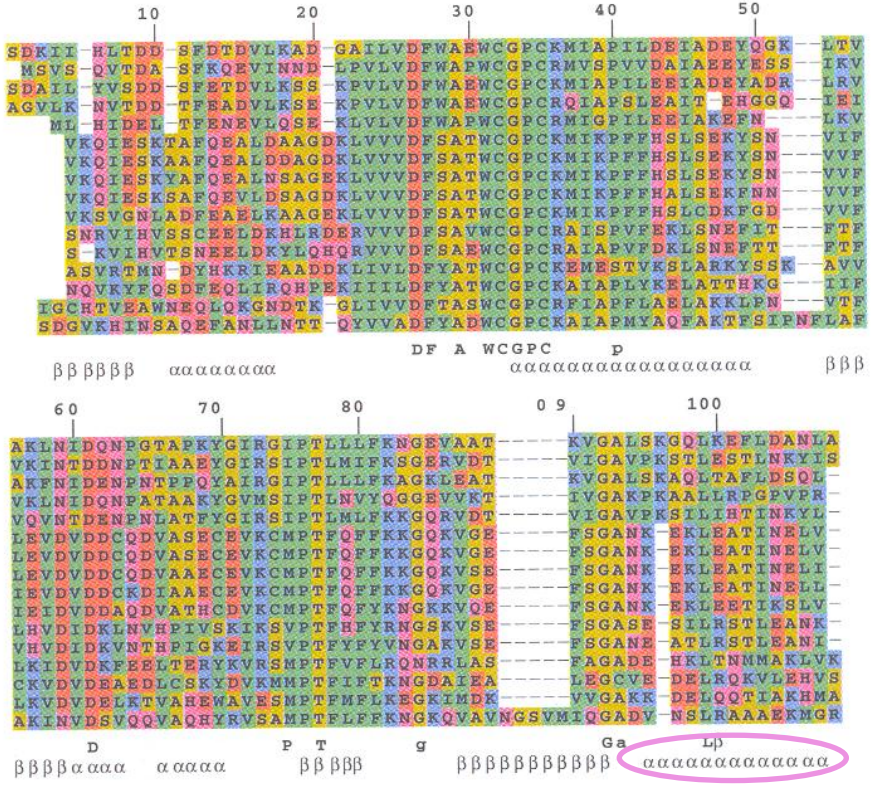
3) A conserved pattern of hydrophobic elements separated by one residue, where the amino acids in between show more variation and can be hydrophiliic suggests a β-sheet on the molecular surface.

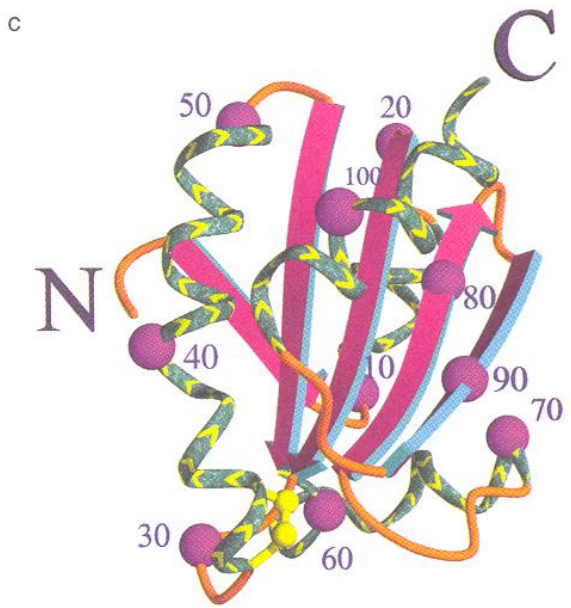




Infos from MSA of Thioredoxin family

4) A conserved pattern of hydrophobic amino acids at distances of about 4 residues suggest an α-helix.





Progressive Multiple Sequence Alignment

- Presented by Feng & Doolittle in 1987
- ist a **heuristic** technique.

  Therefore does not guarantee to find the “optimal” alignment.

- Needs *(n-1) + (n-2) + (n-3) ... (n-n+1)* pairwise sequence alignments at the start.

- Very populart implementation in **Clustal** available (Des Higgins)

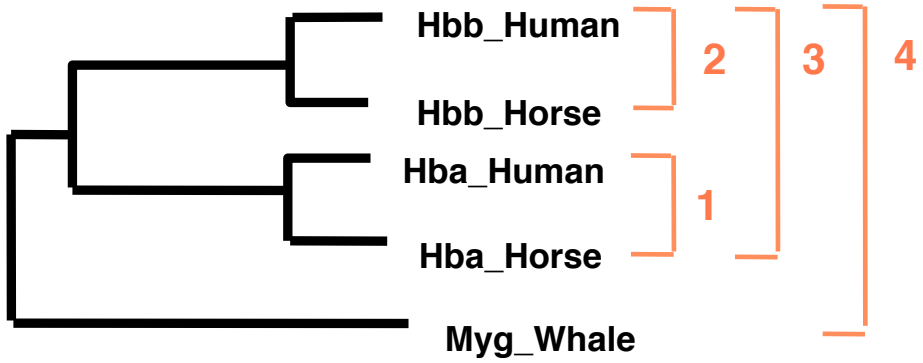
- **ClustalW** is a later version using **weights**.
- The ClustalW is one of the 10-most cited papers in biology (> 30.000).

ClustalW- Pairwise Alignments

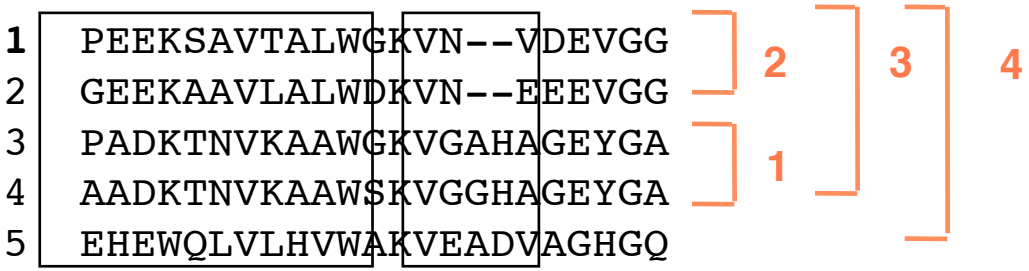
- Compute all *(n-1)+(n-2)...(n-n+1)* pairwise alignments.
- Derive from each pairwise alignments the “distance” between this pair. (one can use e.g. 1 – fractional sequence similarity)
- Construct a **distance matrix**.
- Construct a **neighbor tree** from the pair wise distances (Ruslan, thursday)
- This tree indicates the **sequence** for performing the progressive multiple sequence alignment.

Overview of the ClustalW procedure

```
Hbb_Human   1   -
Hbb_Horse     2   .17   -
Hba_Human   3   .59   .60   -
Hba_Horse     4   .59   .59   .13   -
Myg_Whale    5   .77   .77   .75   .75   -
```



alpha-helices



CLUSTAL W

↓

Fast pairwise alignments:
Set up distance matrix

↓

Neighbor-tree

↓

progressive alignments
according to the tree

ClustalW- pro's and con's

Pro:

- Relative good speed

Con:

- No objective function
- No statistics available to quantify whether alignment is good or bad (see E-value for BLAST)
- No chance to judge whether alignment is "correct"

Possible Problem:

- Procedure can get stuck in "local minimum"

  If a "mistake" was introduced into the alignment at an early time, this cannot be corrected subsequently because relative position of aligned sequences stays fixed

MSA with MAFFT program

Aim: detect **local relatedness** of two sequences (homologous segments) by analyzing their correlation.

This can be done very quickly by **Fast Fourier Transformation**.

Needed is a **numerical representation** of both sequences.

Assume: most important in evolution is the **volume** and **polarity** of each sequence.

Set up 2 vectors of length *n* that contain the volumes and polarities of all *n* amino acids.

MSA with MAFFT program

Compute correlation of 2 vectors v_1 , v_2 containing the amino acids volumes for any relative translation k:

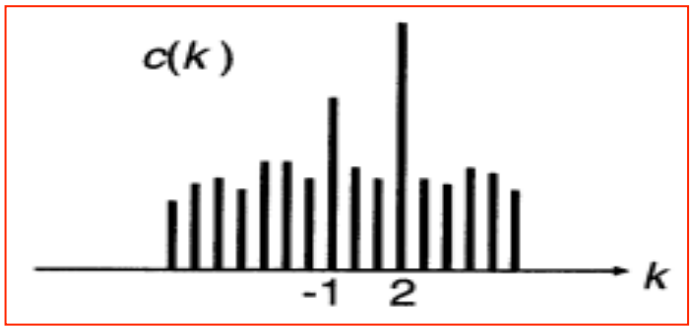
$$c_v(k) = \sum_{1 \le n \le N, 1 \le n+k \le M} \hat{v}_1(n)\hat{v}_2(n+k).$$

In the same way, compute the correlation of the polarity vectors.

Take the sum of both correlations

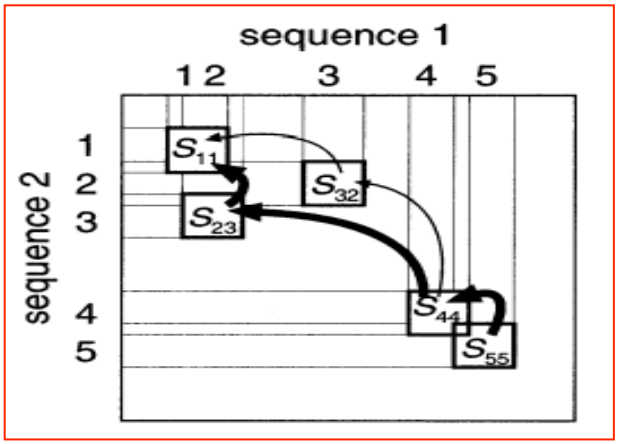
$$c(k) = c_v(k) + c_p(k).$$

Step 1: Find matching segments (i.e. potentially homologous) with maximal correlation



MSA mit MAFFT-Programm

Step 2: Form pairwise alignments with limited global dynamic programming.



Step 3: construct progressively multiple alignment:

- Fast computation of distance matrix:
  - → group 20 amino acids in 6 physico chemical groups
  - → count 6-Tuples common in both sequences (see Blast)
- construct tree with UPGMA method
- Set up multiple alignment according to tree

Step 4: refine MSA iteratively by splitting MSA in 2 regions and re-align them

Alignment of protein coding DNA sequences

- **Aligning DNA sequences coding for proteins makes little sense**
- **Better translate them in protein sequences and the align**

```
ATGCTGTTAGGG                ATGCT-GTTAGGG
ATGCTCGTAGGG      ------>   ATGCTCGT-AGGG
```

Summary

Progressive alignments are most popular for computing MSAs

Multiple sequence alignment ist **not trivial**. Manually correcting the alignments may improve the alignment.

Generating MSAs promotes thinking in **protein families** and –**function**.