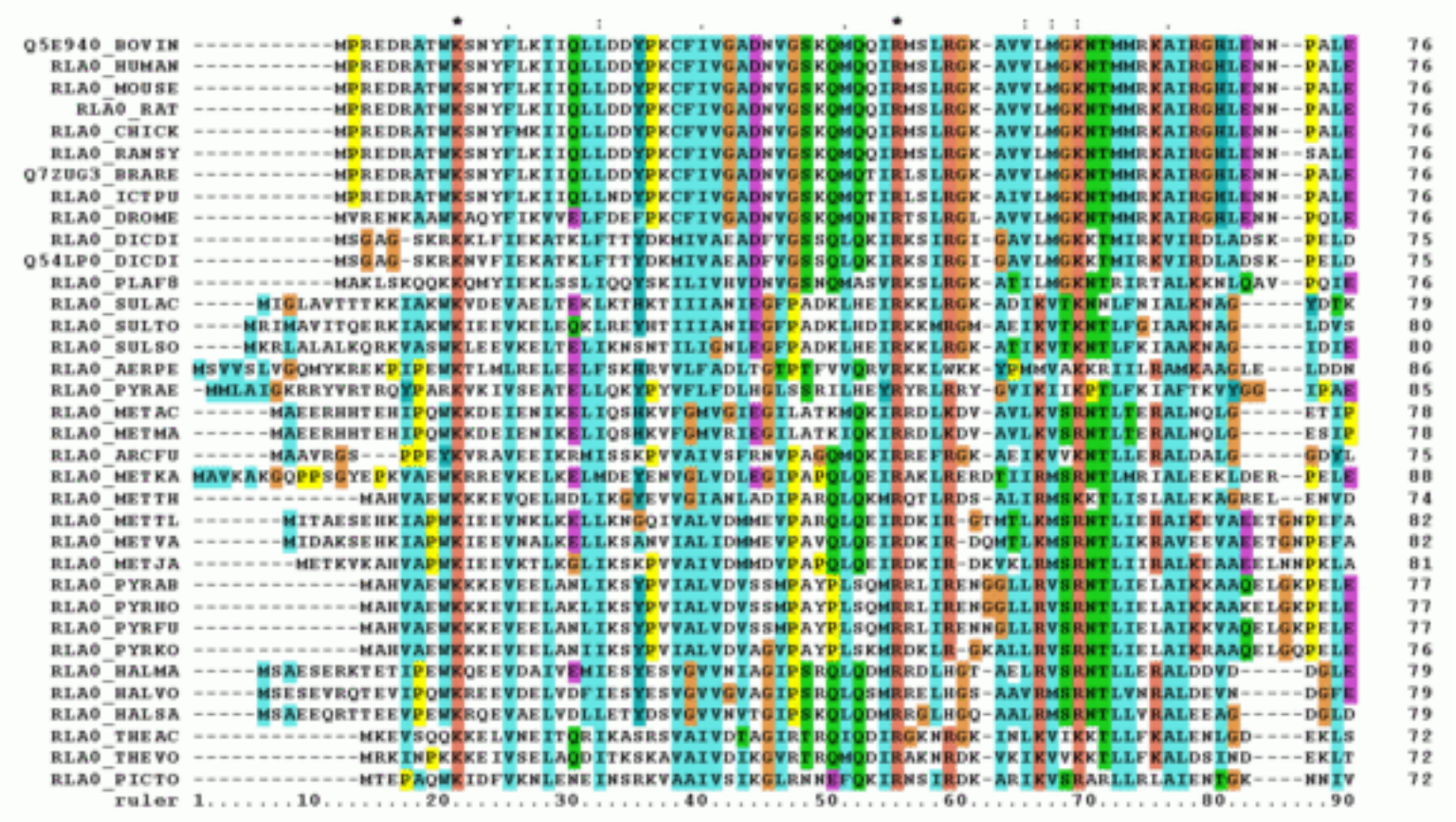
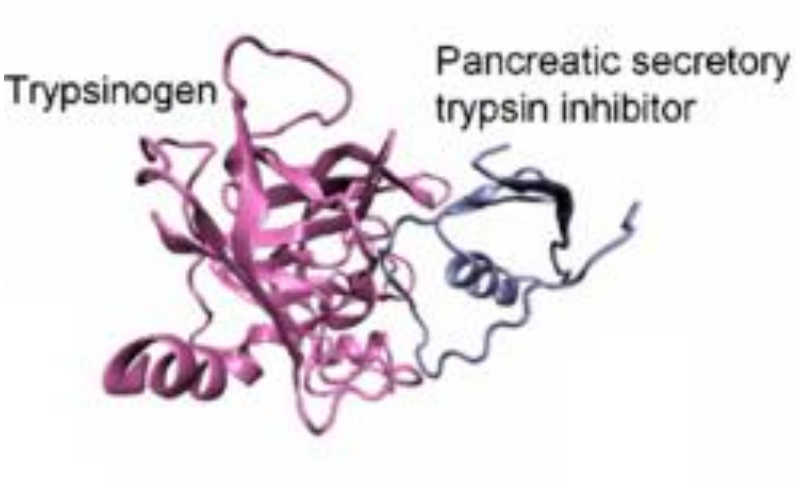
Multiple Sequence Alignment

- a quick overview -



Motivations for sequence alignment

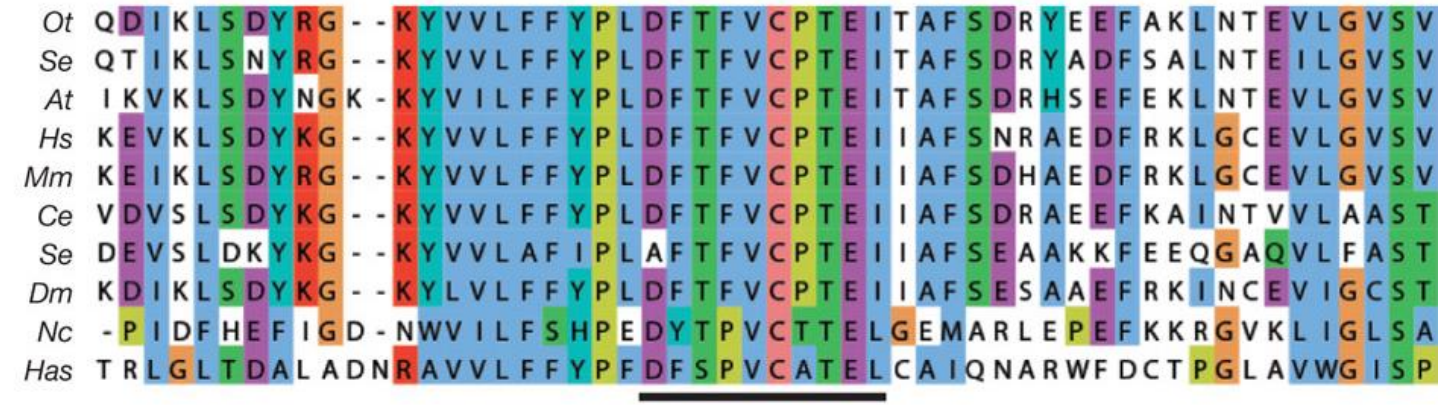
1) find genes that are related by common descent

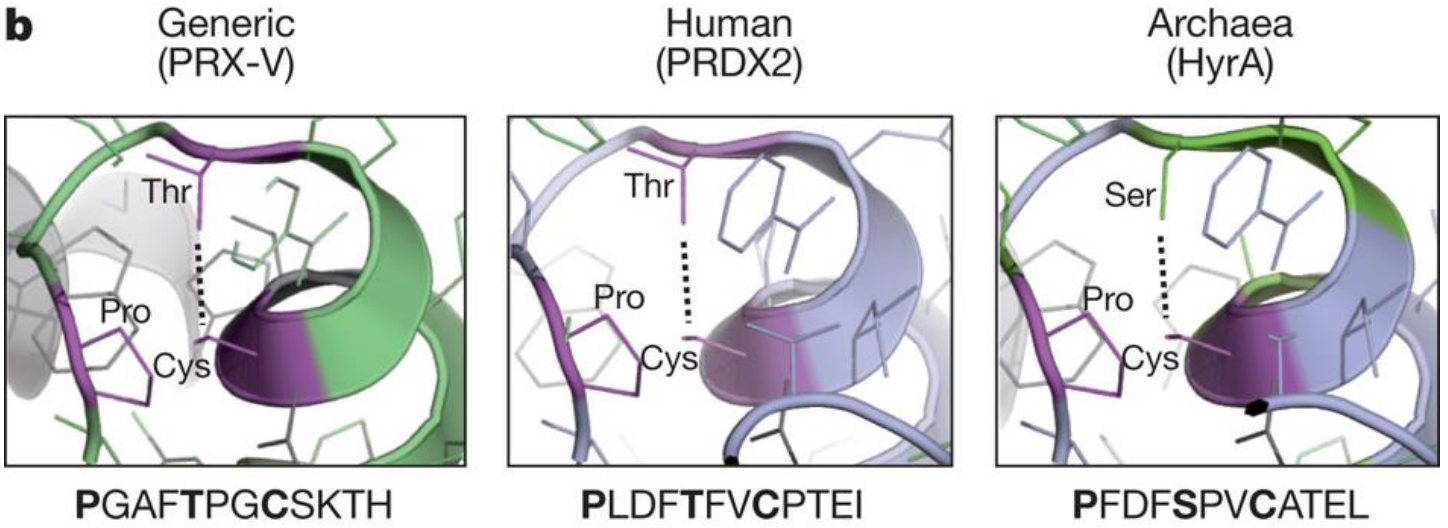



```
Chymotrypsin  VKKTMVCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSFGSRRGC [...]
              +   M C G  +G   +C GDSGGP+ C   NG  +  G+VS+G   GC
Trypsinogen   ITSNMFCVGFLEGGKDSCQGDSGGPVVC---NGQLQ--GVVSWGD--GC [...]
```

Motivations for sequence alignment

2) to identify and check the state of “active sites”



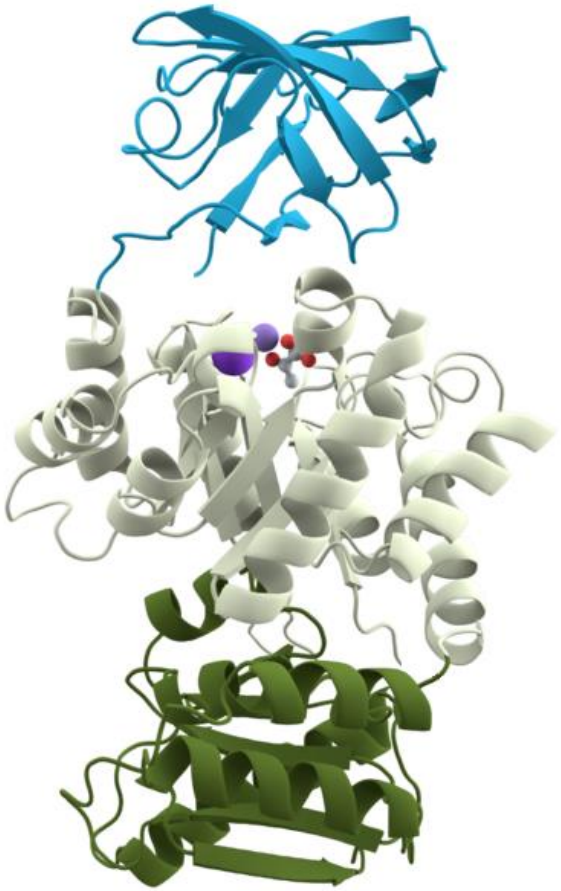


From: Peroxiredoxins are conserved markers of circadian rhythms. Nature 485, 459–464 (24 May 2012)

discover domains by looking at alignments of known domains to find out domains in unknown seqs

Motivations for sequence alignment

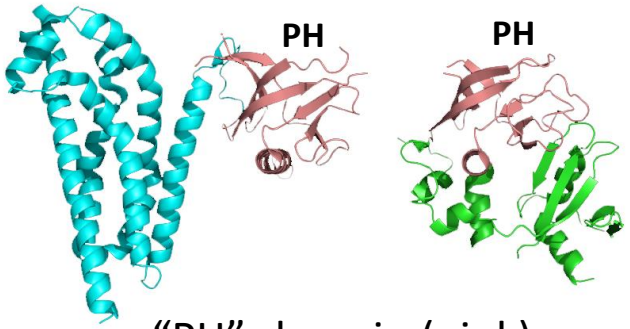
3) to identify and characterize “protein domains”



Pyruvate Kinase

definition:

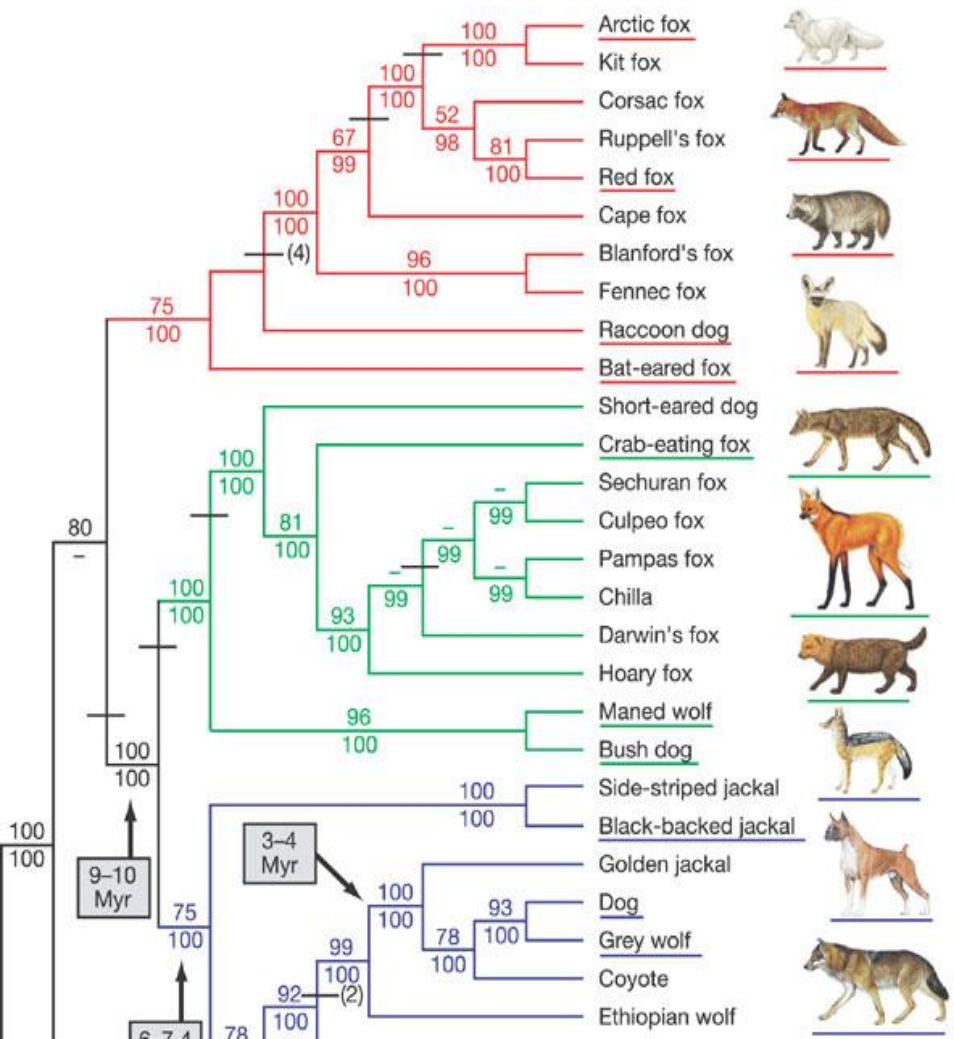
“parts of proteins that can evolve, function, and exist <u>independently</u> of the rest”

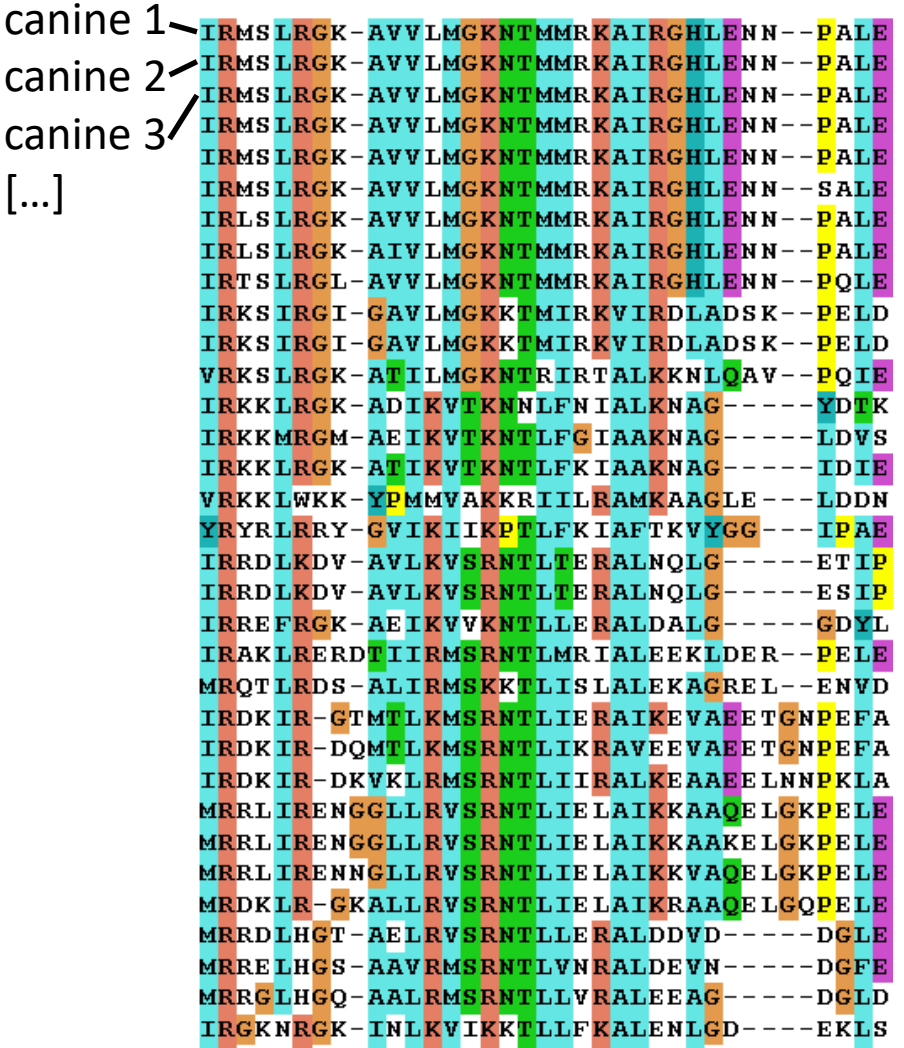


“PH”-domain (pink);
occurring in two different proteins

Motivations for sequence alignment

4) to make phylogenetic inferrences (“trees”)





How it's done …

```
Chymotrypsin   VKKTMVCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSFGSRRGC [...]
               +   M C G  +G    +C GDSGGP+ C   NG  +  G+VS+G   GC
Trypsinogen    ITSNMFCVGFLEGGKDSCQGDSGGPVVC---NGQLQ--GVVSWGD--GC [...]
```

```
NCQLENGSWEV
  C   NG  +
VC---NGQLQ-
```

- Why not align "QL" instead of "NG" ??
- What does the "+" mean ?
- How "good" is my alignment?

=> "substitution matrix" (learned from structures)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Pairwise Alignment

a) BLAST: quick and dirty

```
VKKTMVCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSF
                      GGP
```

1) identify word-matches
(use indexing)

```
VKKTMVCAGG-DGVISACNGDSGGPLNCQLENGSWEVFGIVSF
+    M C G   +G    +C GDSGGP+ C    NG   +   G+VS+
ITSNMFCVGFLEGGKDSCQGDSGGPVVC---NGQLQ--GVVSW
```

2) extend into an "HSP"
(high-scoring sequence pair)

b) Dynamic Programming – correct and slow

	A	T	T	G	G	A	G	T	C	G
A	1	⇨0	⇨-1	⇨-2	⇨-3	⇨-4	⇨-5	⇨-6	⇨-7	⇨-8
T	⇩0	↘2	↘1	→0	→-1	→-2	⇨-3	↘⇨-4	⇨-5	⇨-6
C	⇩-1	⇩1	↘2	→1	→0	→-1	↘→-2	⇨-3	↘-3	⇨-4
T	⇩-2	↘⇩0	↘2	↘2	↘⇨1	↘⇨0	↘⇨-1	↘-1	⇨-2	⇨-3
C	⇩-3	⇩-1	⇩1	↘2	↘2	↘⇨1	↘⇨0	↘⇨-1	↘0	↘⇨-1
G	⇩-4	⇩-2	⇩0	↘2	↘3	↘⇨2	↘2	⇨1	⇨0	↘1

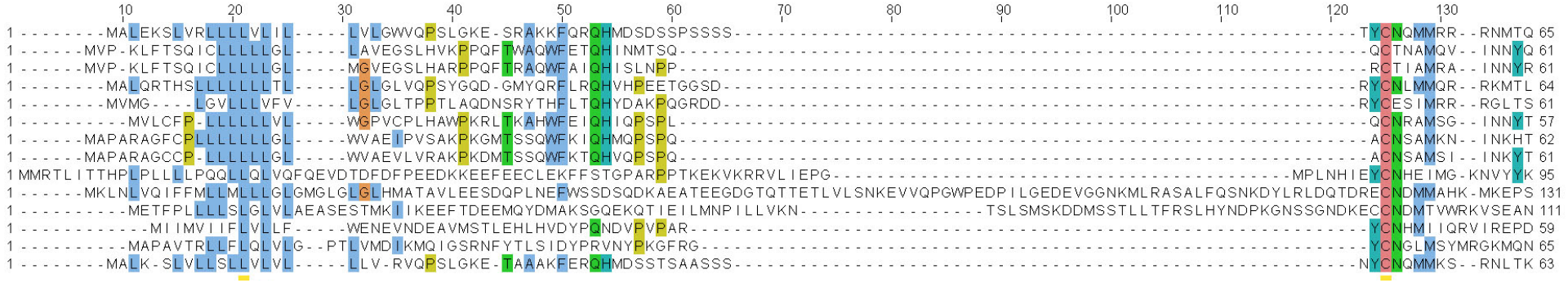
```
ATTGGAGTCG        83%
ATC----TCG

OR

ATTGGAGTCG        83%
AT----CTCG
```

explore all possible paths
(but excluding obvious
dead ends)

Multiple Alignment



Combinatorial Explosion: very many possible solutions

Complexity: O(alignment_length $^{number\_seqs}$)

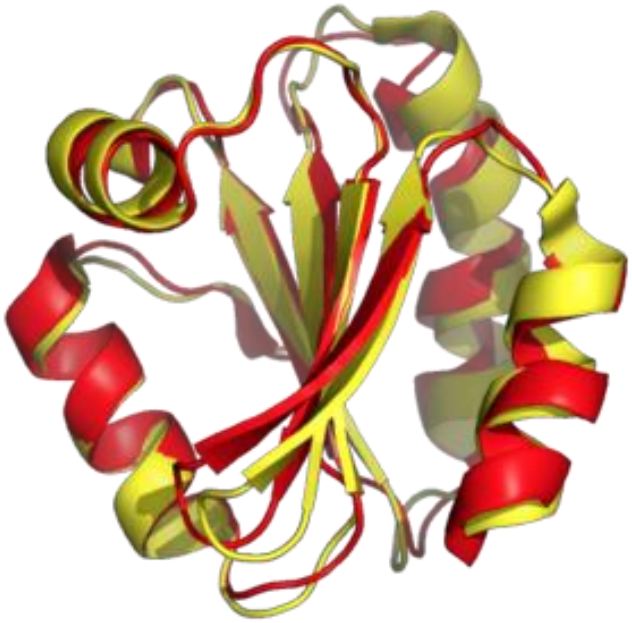
=> an NP-complete problem !!

not feasible

Multiple Alignment

Quality of MSA: Benchmarking

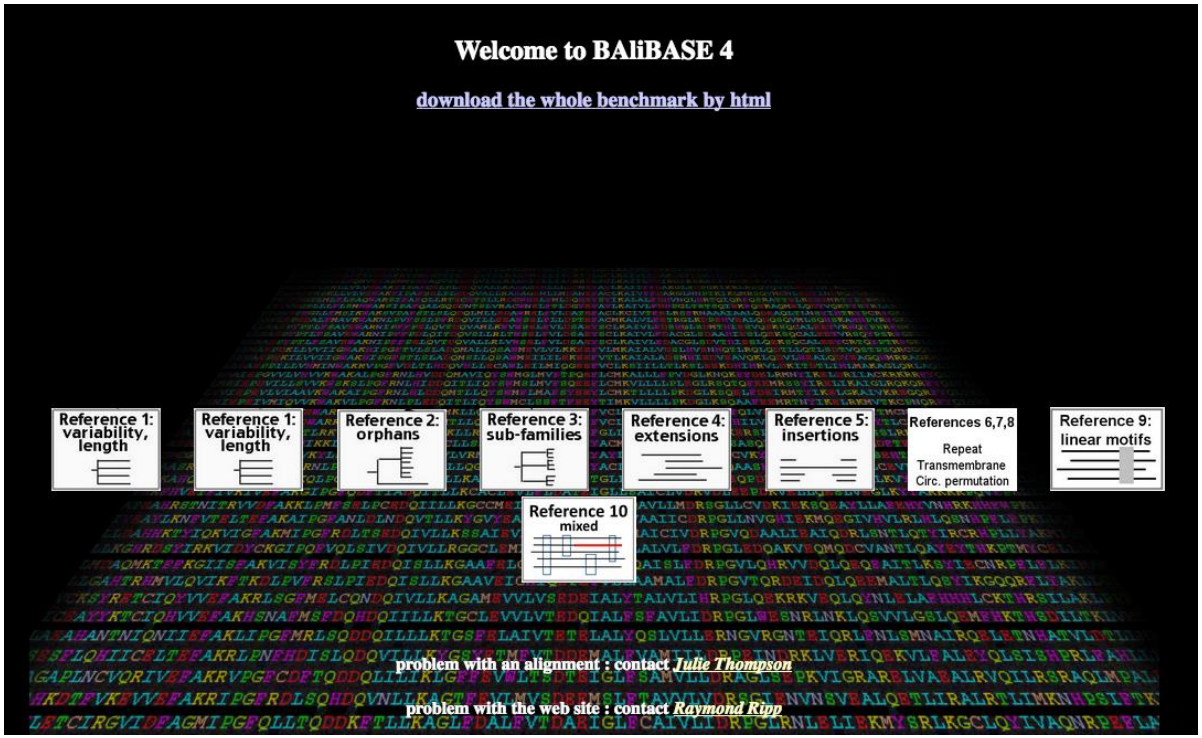
has high quality alignments in this database



Structural Aligments offer the best benchmarks !

"BAliBASE":
Benchmark Alignment Database

Hand-made multiple sequence alignments
Based on selected structural alignments



Practical Session: Source of your Sequences

Dental Calculus of Medieval Mummy

(with kind permission of Christina Warinner, UZH)