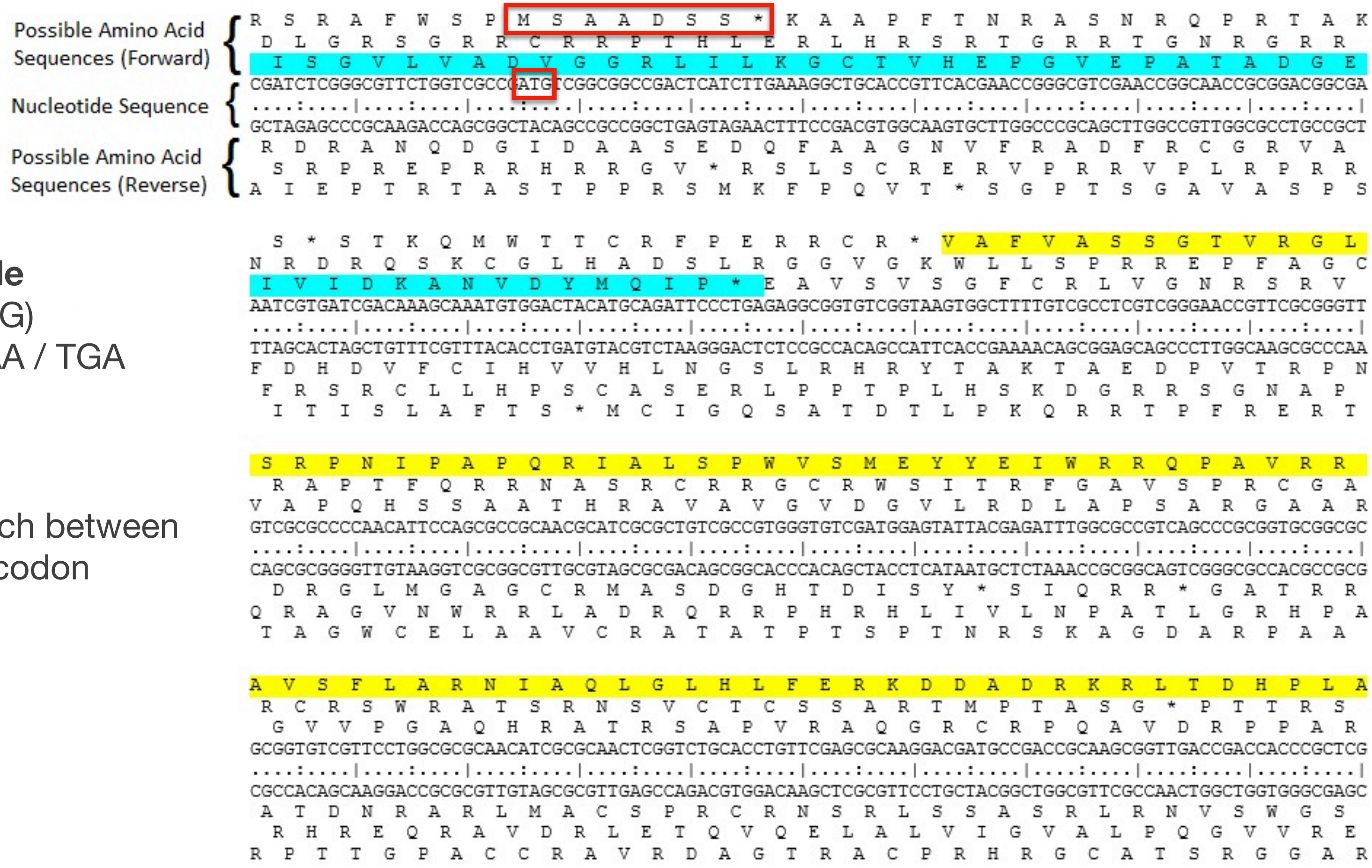
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Help!! What is ....

... an ORF ??

ORF = open reading frame



open reading frame

Possible Amino Acid Sequences (Forward)

Nucleotide Sequence

Possible Amino Acid Sequences (Reverse)

```
R  S  R  A  F  W  S  P  M  S  A  A  D  S  S  *  K  A  A  P  F  T  N  R  A  S  N  R  Q  P  R  T  A  K
 D  L  G  R  S  G  R  R  C  R  R  P  T  H  L  E  R  L  H  R  S  R  T  G  R  R  T  G  N  R  G  R  R
  I  S  G  V  L  V  A  D  V  G  G  R  L  I  L  K  G  C  T  V  H  E  P  G  V  E  P  A  T  A  D  G  E
CGATCTCGGGCGTTCTGGTCGCCGATGTCGGCGGCCGACTCATCTTGAAAGGCTGCACCGTTCACGAACCGGGCGTCGAACCGGCAACCGCGGACGGCGA
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
GCTAGAGCCCGCAAGACCAGCGGCTACAGCCGCCGGCTGAGTAGAACTTTCCGACGTGGCAAGTGCTTGGCCCGCAGCTTGGCCGTTGGCGCCTGCCGCT
 R  D  R  A  N  Q  D  G  I  D  A  A  S  E  D  Q  F  A  A  G  N  V  F  R  A  D  F  R  C  G  R  V  A
  S  R  P  R  E  P  R  R  H  R  R  G  V  *  R  S  L  S  C  R  E  R  V  P  R  R  V  P  L  R  P  R  R
A  I  E  P  T  R  T  A  S  T  P  P  R  S  M  K  F  P  Q  V  T  *  S  G  P  T  S  G  A  V  A  S  P  S

  S  *  S  T  K  Q  M  W  T  T  C  R  F  P  E  R  R  C  R  *  V  A  F  V  A  S  S  G  T  V  R  G  L
N  R  D  R  Q  S  K  C  G  L  H  A  D  S  L  R  G  G  V  G  K  W  L  L  S  P  R  R  E  P  F  A  G  C
 I  V  I  D  K  A  N  V  D  Y  M  Q  I  P  *  E  A  V  S  V  S  G  F  C  R  L  V  G  N  R  S  R  V
AATCGTGATCGACAAAGCAAATGTGGACTACATGCAGATTCCCTGAGAGGCGGTGTCGGTAAGTGGCTTTTGTCGCCTCGTCGGGAACCGTTCGCGGGTT
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
TTAGCACTAGCTGTTTCGTTTACACCTGATGTACGTCTAAGGGACTCTCCGCCACAGCCATTCACCGAAAACAGCGGAGCAGCCCTTGGCAAGCGCCCAA
F  D  H  D  V  F  C  I  H  V  V  H  L  N  G  S  L  R  H  R  Y  T  A  K  T  A  E  D  P  V  T  R  P  N
 F  R  S  R  C  L  L  H  P  S  C  A  S  E  R  L  P  P  T  P  L  H  S  K  D  G  R  R  S  G  N  A  P
  I  T  I  S  L  A  F  T  S  *  M  C  I  G  Q  S  A  T  D  T  L  P  K  Q  R  R  T  P  F  R  E  R  T

 S  R  P  N  I  P  A  P  Q  R  I  A  L  S  P  W  V  S  M  E  Y  Y  E  I  W  R  R  Q  P  A  V  R  R
  R  A  P  T  F  Q  R  R  N  A  S  R  C  R  R  G  C  R  W  S  I  T  R  F  G  A  V  S  P  R  C  G  A
V  A  P  Q  H  S  S  A  A  T  H  R  A  V  A  V  G  V  D  G  V  L  R  D  L  A  P  S  A  R  G  A  A  R
GTCGCGCCCCAACATTCCAGCGCCGCAACGCATCGCGCTGTCGCCGTGGGTGTCGATGGAGTATTACGAGATTTGGCGCCGTCAGCCCGCGGTGCGGCGC
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CAGCGCGGGGTTGTAAGGTCGCGGCGTTGCGTAGCGCGACAGCGGCACCCACAGCTACCTCATAATGCTCTAAACCGCGGCAGTCGGGCGCCACGCCGCG
  D  R  G  L  M  G  A  G  C  R  M  A  S  D  G  H  T  D  I  S  Y  *  S  I  Q  R  R  *  G  A  T  R  R
Q  R  A  G  V  N  W  R  R  L  A  D  R  Q  R  R  P  H  R  H  L  I  V  L  N  P  A  T  L  G  R  H  P  A
 T  A  G  W  C  E  L  A  A  V  C  R  A  T  A  T  P  T  S  P  T  N  R  S  K  A  G  D  A  R  P  A  A

A  V  S  F  L  A  R  N  I  A  Q  L  G  L  H  L  F  E  R  K  D  D  A  D  R  K  R  L  T  D  H  P  L  A
 R  C  R  S  W  R  A  T  S  R  N  S  V  C  T  C  S  S  A  R  T  M  P  T  A  S  G  *  P  T  T  R  S
  G  V  V  P  G  A  Q  H  R  A  T  R  S  A  P  V  R  A  Q  G  R  C  R  P  Q  A  V  D  R  P  P  A  R
GCGGTGTCGTTCCTGGCGCGCAACATCGCGCAACTCGGTCTGCACCTGTTCGAGCGCAAGGACGATGCCGACCGCAAGCGGTTGACCGACCACCCGCTCG
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CGCCACAGCAAGGACCGCGCGTTGTAGCGCGTTGAGCCAGACGTGGACAAGCTCGCGTTCCTGCTACGGCTGGCGTTCGCCAACTGGCTGGTGGGCGAGC
 A  T  D  N  R  A  R  L  M  A  C  S  P  R  C  R  N  S  R  L  S  S  A  S  R  L  R  N  V  S  W  G  S
  R  H  R  E  Q  R  A  V  D  R  L  E  T  Q  V  Q  E  L  A  L  V  I  G  V  A  L  P  Q  G  V  V  R  E
R  P  T  T  G  P  A  C  C  R  A  V  R  D  A  G  T  R  A  C  P  R  H  R  G  C  A  T  S  R  G  G  A  R
```

Standard genetic code
start codon = ATG (AUG)
stop codon = TAG / TAA / TGA

ORF = sequence stretch between start codon and stop codon

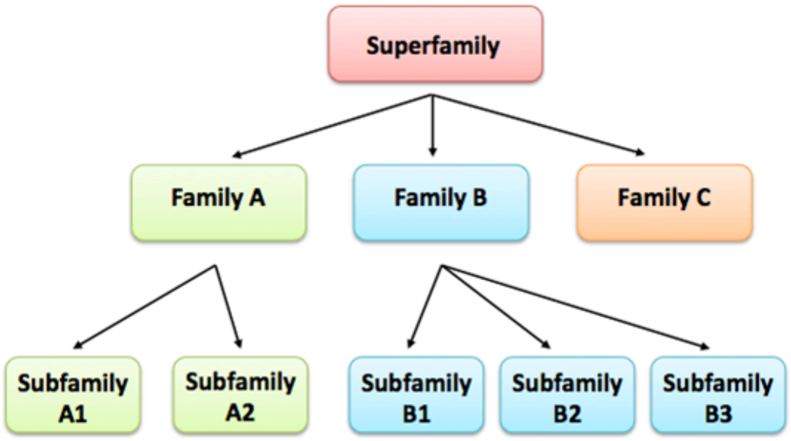
Help!! What is ....

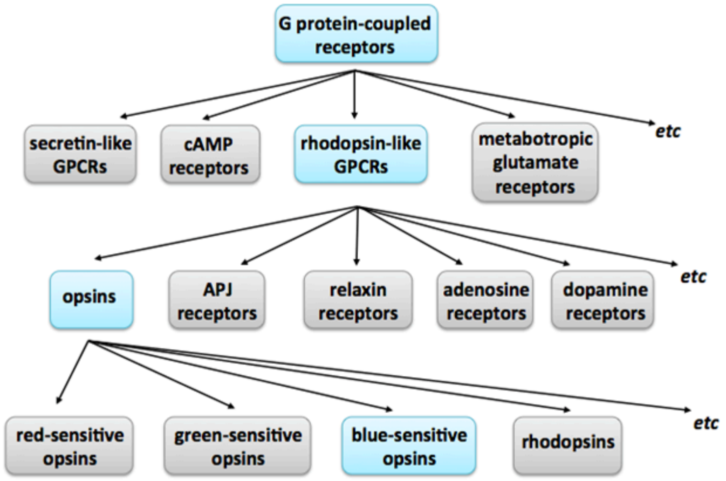
... a protein (super)family ??

Families, superfamilies,...

- **Protein family** = group of proteins that share a **common evolutionary origin**, reflected by their related functions and similarities in sequence or structure.
    - superfamily = large group of distantly related proteins
    - subfamily = small group of closely related proteins

- **Protein families** are often arranged into hierarchies, with proteins that share a common ancestor subdivided into smaller, more closely related groups.

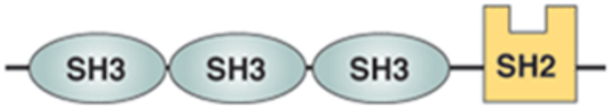


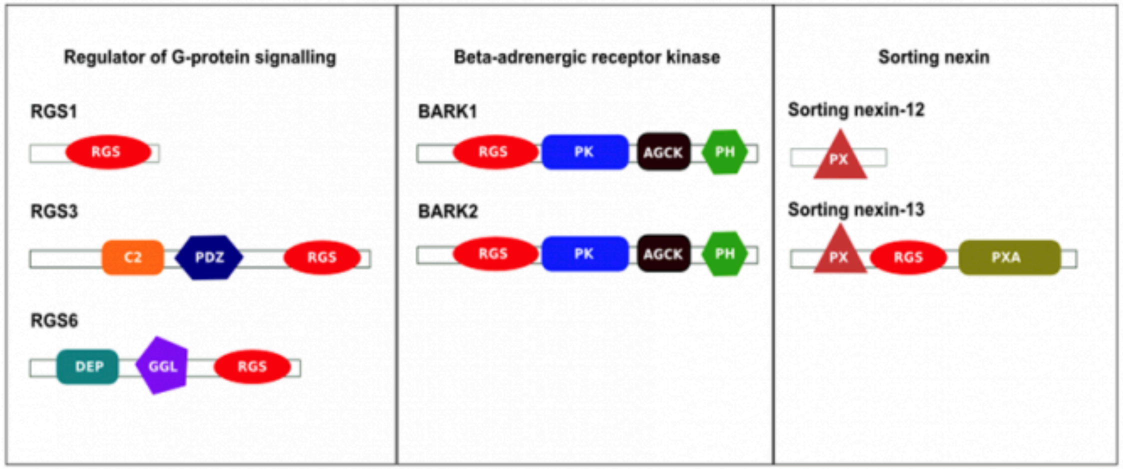


Protein domains

- Protein domains = **functional** and/or **structural** units in a protein
- Protein usually contain **several protein domains**

SH3 domain





RGS family

beta-adrenergic receptor kinase family

Sorting nexin family

RGS = Regulator of G-protein signalling

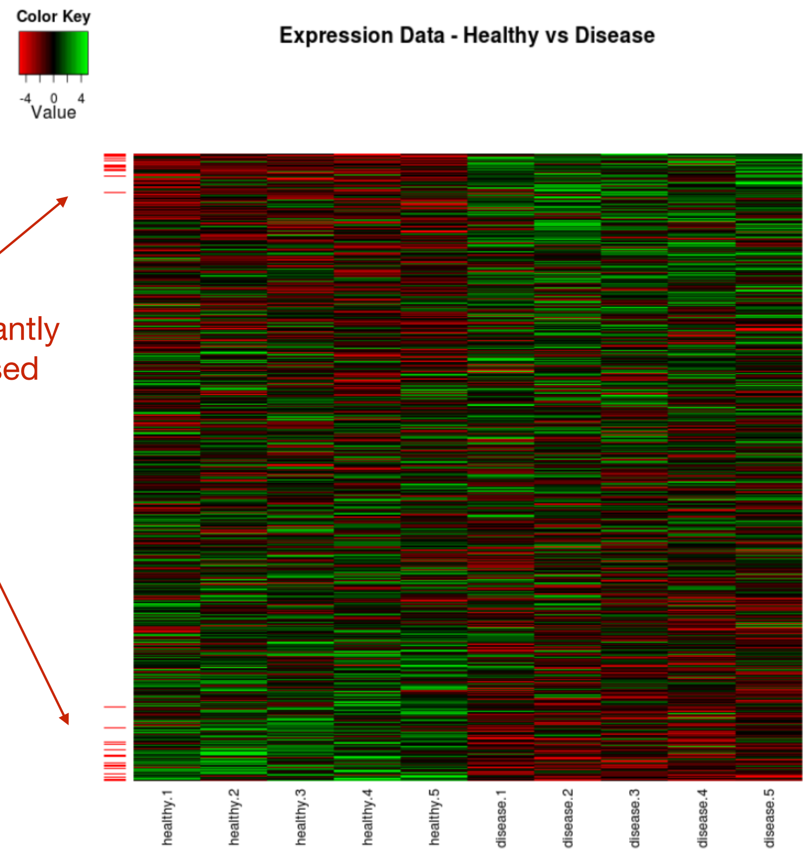
Help!! What is ....

... an E-value ??

Fake gene expression data

- Finding differentially expressed genes between healthy and disease patients
- t-test with α = 5%
- H_0: non-significant expression difference between the two groups

This dataset contains only random numbers
→ H_0 holds for all 10.000 "genes"
→ all the 445 genes are <u>false-positives</u>



445 significantly diff. expressed genes

Multiple testing

- **Significance level** α: level at which to **reject** $(p < \alpha)$ or **accept** $(p > \alpha)$ the Null hypothesis
- **P-value**: probability to observe a more extreme effect if H0 is true ("risk of a false-positive by random chance")
- **E-value**: expected number of false-positive events when N tests are performed

$$E = p \cdot N$$

Help!! What is ....

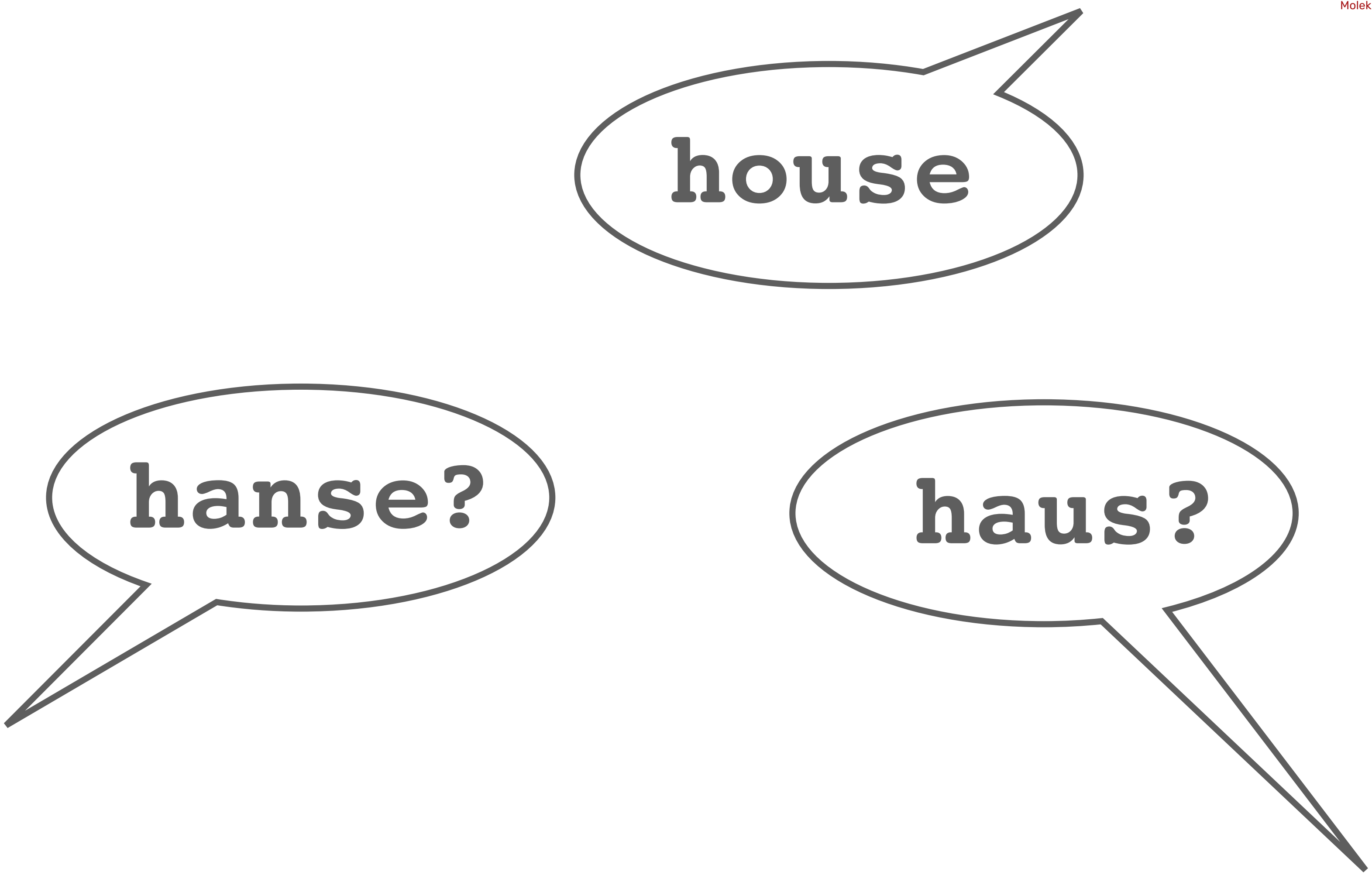
... BLAST ??

Why sequence alignments ?

```
>Protein sequence
MLCPISGWAIYSKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSN
GTVKDRSPYRTLMSCPVGEAPSPYNSRFESVAWSASACHDGISWLTIGISGPDNGAV
AVLKYNGIITDTIKSWRNNTLRTQESECACVNGSCFTVMTDGPSNEQASYKIFKIEK
```

- Open questions
  - **Homologues**: *are there related sequences in other organisms?*
  - **Function**: *possible biological/molecular/enzymatic function?*
  - **Origin**: *from which organisms / clade? (Example: Metagenomics)*
- **Global** (= whole sequence) / **local** (= parts of the sequence) comparisons

Lost in translation...



house

hanse?

haus?

Lost in translation...

house	house
hanse	haus

Lost in translation...

```
house     house
|  ||     | ||
hanse     haus
```

3 Matches

2 Mismatches

3 Matches

1 Mismatch

1 Ins./Del. (Indel)

Lost in translation...

```
house
|  ||
hanse
```

```
house
| ||
haus
```

3 Matches

2 Mismatches

3 Matches

1 Mismatch

1 Ins./Del. (Indel)

- which comparison is better?
  → ***Scoring-Method***
- **Score** should take into account...
  - Matches (+)
  - Mismatches (-)
    - Vokal/Vokal oder Kons./Kons. (-)
    - Vokal/Kons. (--)
  - Insertion/Deletion (-)

Lost in translation...

```
house     house
|  ||     |  ||
hanse     haus
```

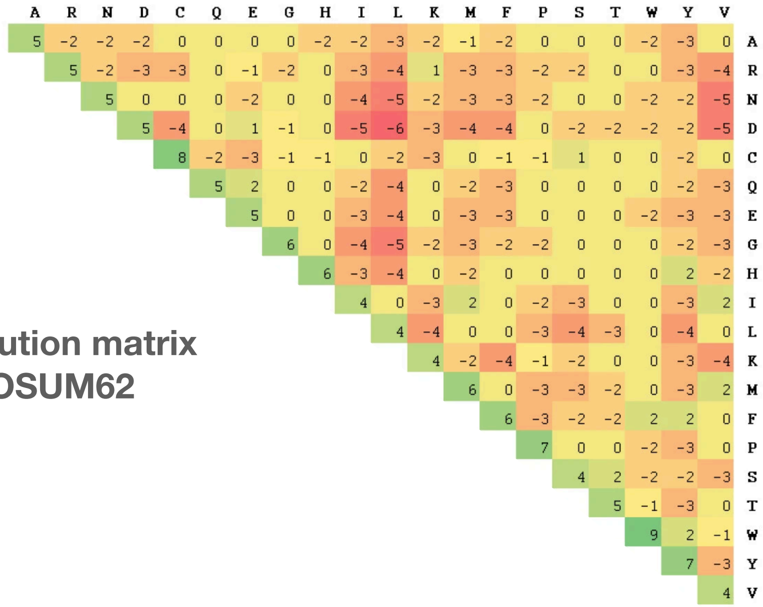
3 x (+1) - 1 x (-1)
-1 x (-2)
= **0**

3 x (+1) - 1 x (-1)
-1 x (-1)
= **+1**

- which comparison is better?
  → ***Scoring-Method***
- **Score** should take into account...
  - Matches (+)
  - Mismatches (-)
    - Vokal/Vokal oder Kons./Kons. (-)
    - Vokal/Kons. (--)
  - Insertion/Deletion (-)

Scoring Verfahren

- **Matches/mismatches**
  subtitution matrix: values represent frequencies of observed substitutions in homologous sequences (ex. BLOSUM62)



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

Substitution matrix BLOSUM62

- Gaps
  mostly affine cost
  - gap opening (O)
  - gap extensions (E)

```
AITP-------VPQ
AVTPQSLPCSSLQQ
```

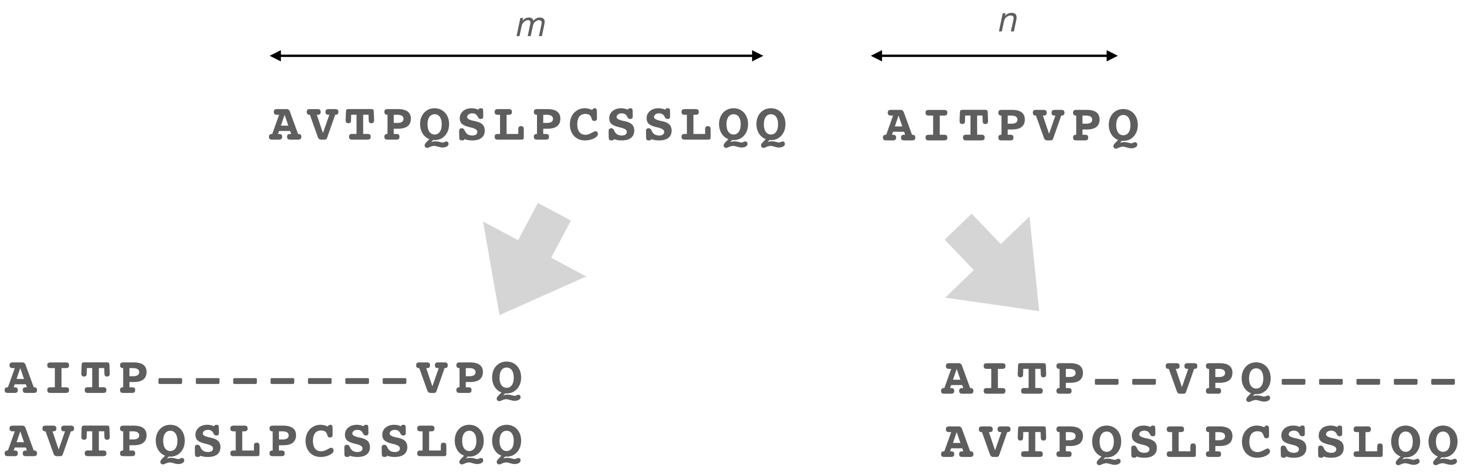
indels l = 7

- opening: O = -11
- extension: E = -2

here: -11 + 7 x (-2) = -25

Which is the best alignment?

- Given 2 sequences, which is the best alignment?



m *n*

```
AVTPQSLPCSSLQQ   AITPVPQ
```

```
AITP--------VPQ          AITP--VPQ-----
AVTPQSLPCSSLQQ           AVTPQSLPCSSLQQ
```

BLOSUM62 Matrix
O = -11; E = -2

$S = -1$ $S = -12$

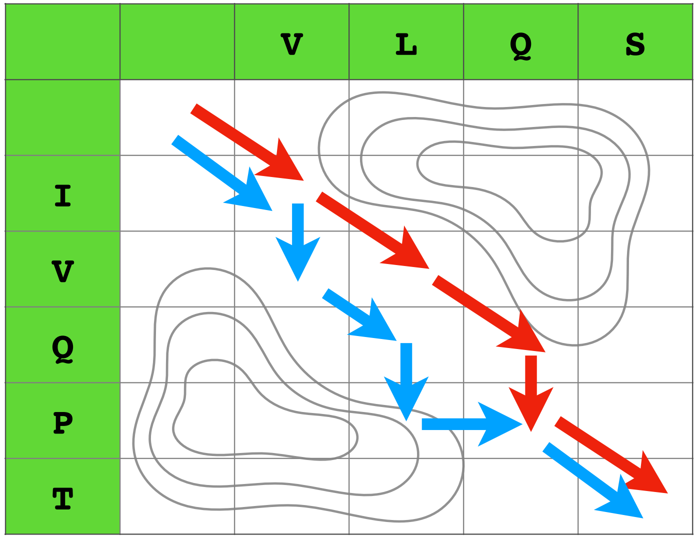
$$\binom{m+n}{n} = \binom{21}{7} =$$ 116.280 possible alignments...

Dynamical programming (DP)

- Dynamical programming allows to determine exactly the best alignment
- Alignment = path in the score matrix
- Best alignment is obtained by determinung at each step the best alignment
- Needleman & Wunsch = **global** alignment
  Smith & Watermann = **local** Alignment

Complexity

$$\mathcal{O}(m \cdot n)$$



V-L-QS
IVQP-T

VLQ-S
IVQPT

→ gap in 2nd sequence

↓ gap in 1st sequence

↘ (mis)match

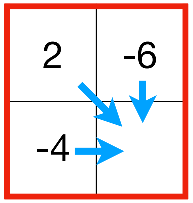
Needlman & Wunsch: globales Alignment

1. Phase

Füllen der Matrix

Gap = -8

		V	L	Q	S
I					
V					
Q					
P					
T					



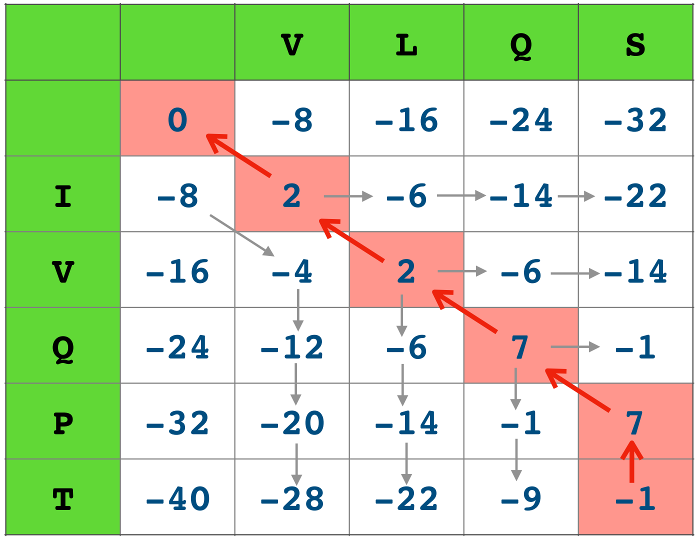
↓ -6-8 = -14

→ -4-8 = -12

↘ **2-0 = 2**

2. Phase

Backtracking von unten rechts
nach oben links



		V	L	Q	S
	0	-8	-16	-24	-32
I	-8	2	-6	-14	-22
V	-16	-4	2	-6	-14
Q	-24	-12	-6	7	-1
P	-32	-20	-14	-1	7
T	-40	-28	-22	-9	-1

```
V L Q S –
I V Q P T
```

Smith Watermann: lokales Alignment

1. Phase

Füllen der Matrix (negative Werte werden durch Null ersetzt)

Gap = -8

		V	L	Q	S
	0	0	0	0	0
I	0	2	0	0	0
V	0	4	2	0	0
Q	0	0	0	7	0
P	0	0	0	0	7
T	0	0	0	0	2

2. Phase

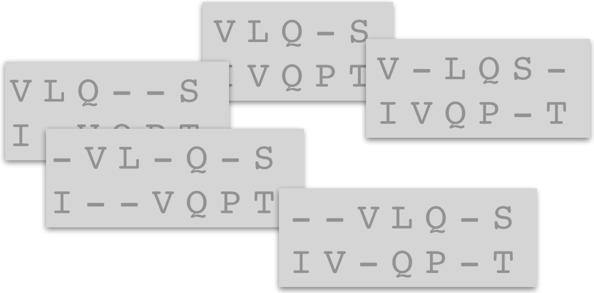
Backtracking vom **höchsten Wert** bis zur ersten Null

		V	L	Q	S
	0	0	0	0	0
I	0	2	0	0	0
V	0	4	2	0	0
Q	0	0	0	7	0
P	0	0	0	0	7
T	0	0	0	0	2

V L Q S
I V Q P

Problem solved?

Enumeration of all possible alignments

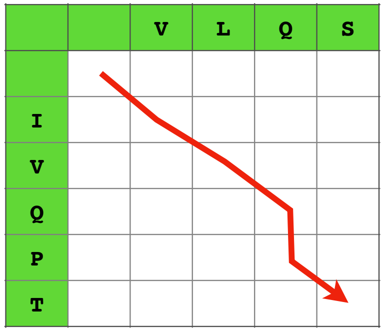


```
V L Q – S
I V Q P T

V – L Q S –
I V Q P – T

V L Q – – S
I

– V L – Q – S
I – – V Q P T

– – V L Q – S
I V – Q P – T
```

complexity
n= m = 50

$$\binom{m+n}{n} = 10^{29}$$

Optimal alignment using DP



		V	L	Q	S
I					
V					
Q					
P					
T					

$$m \cdot n = 2500$$

Optimal alignment with database

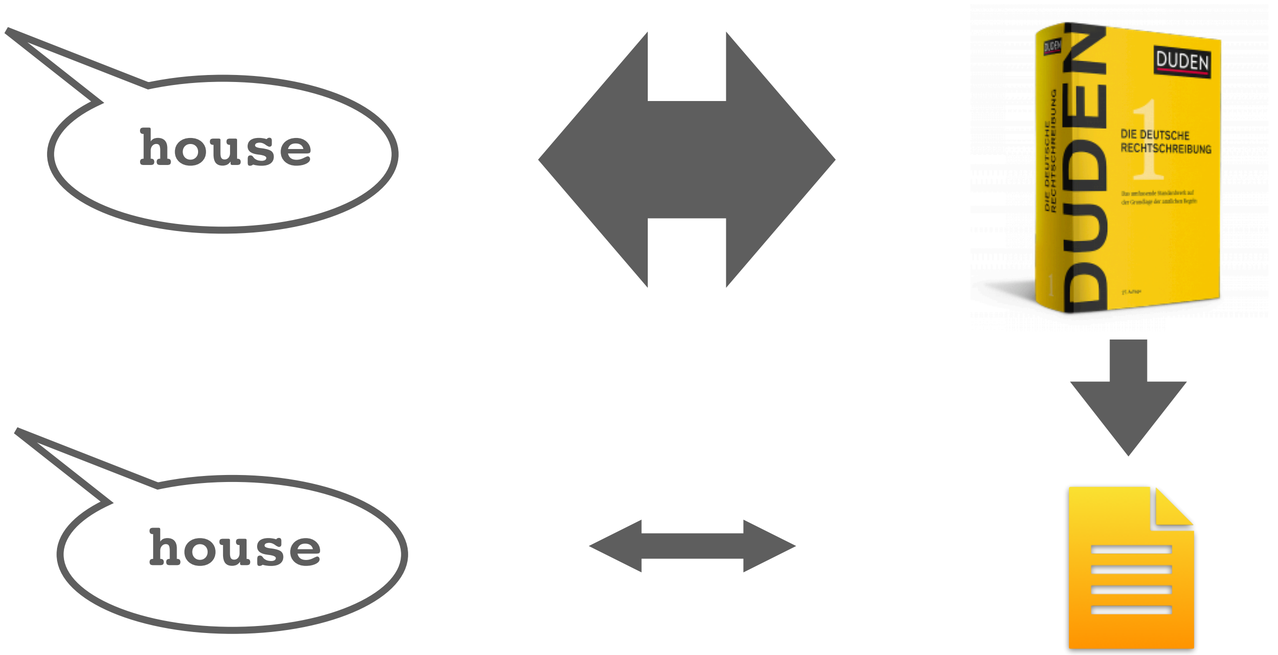
UniProt

N ~ 230 million sequences

$$N \cdot m \cdot n = 5.8 \cdot 10^{11}$$

- **Problem**: DP alignment cannot be computed for all target sequences (too long!)
- **Solution**: select most promissing sequences first... then do DP

Heuristic = short-cut



N ~150.000 words

all words starting with h

N ~1.000 words

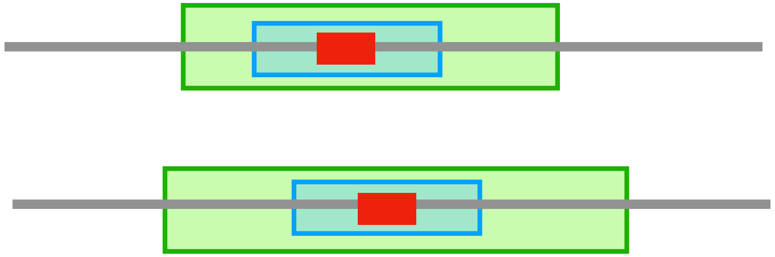
- **Advantage**: much faster!
- **Disadvantage**: maybe the right translation starts with a different letter ...

BLAST: basic local alignment search tool

Heuristic

1. homologous sequences share **very similar short words**
   (Protein: *k*=3; DNA: *k*=11)
2. these words reside in **longer homologous sequences without gaps**
   (HSP = high scoring pairs)
3. starting from HSP **longer alignments with gaps** can be obtained using DP.

→ final raw score S depends on the subsitution matrix, number of matches / mismatches / gaps



BLAST 1.0 (Altschul et al., 1990)
Alignments **ohne** Gaps

BLAST 2.0 (Altschul et al., 1997)
Alignments **mit** Gaps

Score

Raw score obtained from the dynamical programming

Parameter (= black magic)

$$S' = \frac{\lambda S - \ln K}{\ln 2} \text{ (bits)}$$

Size of the database

E = expected number of false-positive in a database of the same size

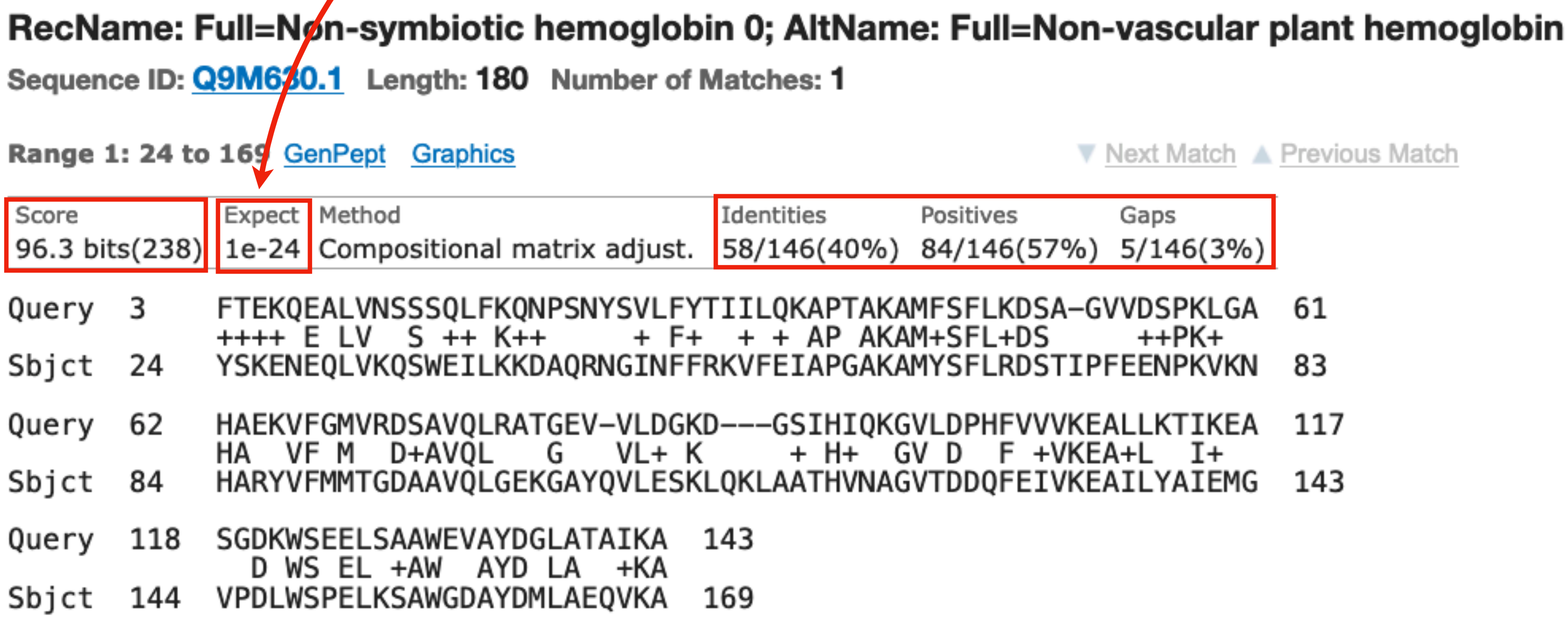
$$E = \frac{Q}{2^{S'}}$$

NCBI BLAST

E-value = number of false-positives with equal scores

$S = 238$

$S' = 96.3$ bits



RecName: Full=Non-symbiotic hemoglobin 0; AltName: Full=Non-vascular plant hemoglobin

Sequence ID: Q9M630.1 **Length:** 180 **Number of Matches:** 1

Range 1: 24 to 169 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
96.3 bits(238)	1e-24	Compositional matrix adjust.	58/146(40%)	84/146(57%)	5/146(3%)

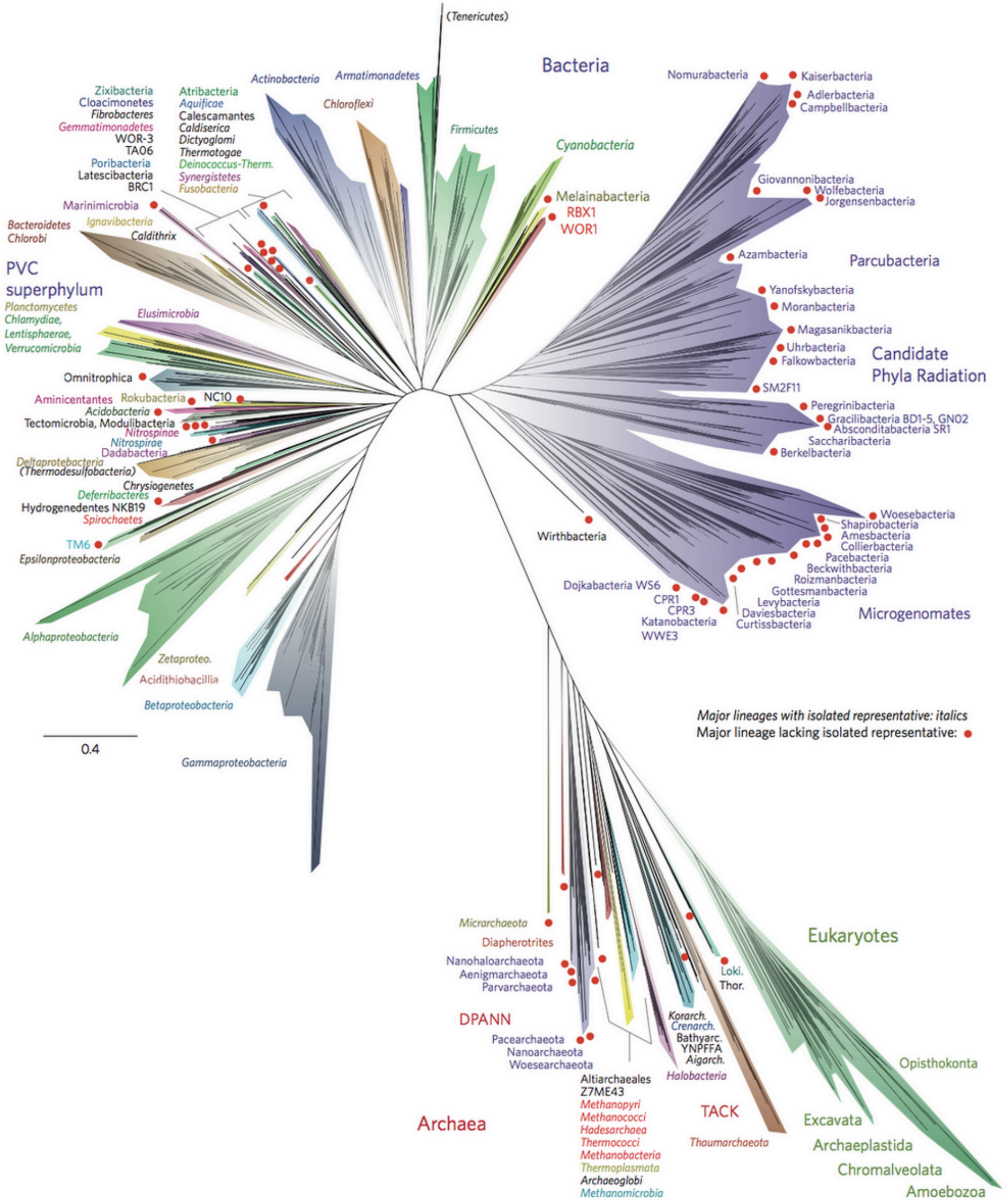
```
Query  3    FTEKQEALVNSSSQLFKQNPSNYSVLFYTIILQKAPTAKAMFSFLKDSA-GVVDSPKLGA  61
            ++++ E LV  S ++ K++     + F+  + + AP AKAM+SFL+DS     ++PK+
Sbjct  24   YSKENEQLVKQSWEILKKDAQRNGINFFRKVFEIAPGAKAMYSFLRDSTIPFEENPKVKN  83

Query  62   HAEKVFGMVRDSAVQLRATGEV-VLDGKD---GSIHIQKGVLDPHFVVVKEALLKTIKEA  117
            HA  VF M  D+AVQL   G   VL+ K     + H+  GV D  F +VKEA+L  I+
Sbjct  84   HARYVFMMTGDAAVQLGEKGAYQVLESKLQKLAATHVNAGVTDDQFEIVKEAILYAIEMG  143

Query  118  SGDKWSEELSAAWEVAYDGLATAIKA  143
              D WS EL +AW  AYD LA  +KA
Sbjct  144  VPDLWSPELKSAWGDAYDMLAEQVKA  169
```

Help!! What is ….

… phylogenetic tree ??

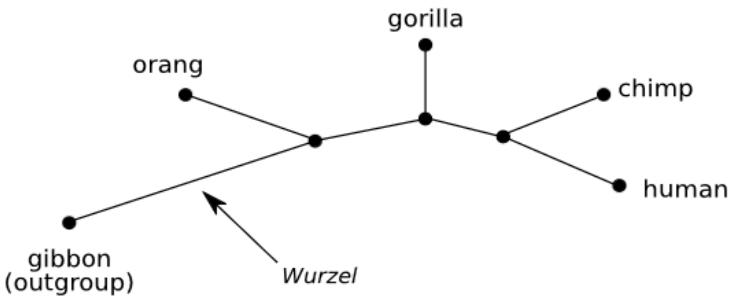
Phylogenetic tree

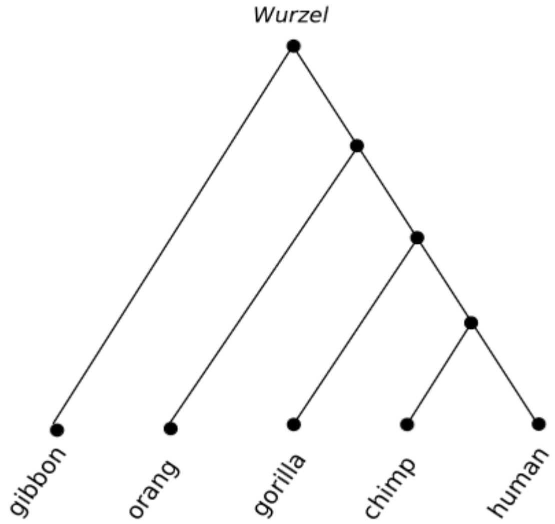
- Represent the evolutionary history of a set of sequences or organisms
- Beware the a tree built from a single gene can differ from the evolutionnary tree of the species!
- Trees are constructed based on multiple alignments, from which a distance matrix can be built


(Tenericutes)
Bacteria
Actinobacteria
Armatimonadetes
Chloroflexi
Firmicutes
Cyanobacteria
Melainabacteria
RBX1
WOR1
Zixibacteria
Cloacimonetes
Fibrobacteres
Gemmatimonadetes
WOR-3
TA06
Poribacteria
Latescibacteria
BRC1
Atribacteria
Aquificae
Calescamantes
Caldiserica
Dictyoglomi
Thermotogae
Deinococcus-Therm.
Synergistetes
Fusobacteria
Marinimicrobia
Bacteroidetes
Chlorobi
Ignavibacteria
Caldithrix
PVC
superphylum
Planctomycetes
Chlamydiae,
Lentisphaerae,
Verrucomicrobia
Elusimicrobia
Omnitrophica
Aminicentantes
Rokubacteria
NC10
Acidobacteria
Tectomicrobia, Modulibacteria
Nitrospinae
Nitrospirae
Dadabacteria
Deltaprotebacteria
(Thermodesulfobacteria)
Chrysiogenetes
Deferribacteres
Hydrogenedentes NKB19
Spirochaetes
TM6
Epsilonproteobacteria
Alphaproteobacteria
Zetaproteo.
Acidithiobacillia
Betaproteobacteria
Gammaproteobacteria
0.4
Nomurabacteria
Kaiserbacteria
Adlerbacteria
Campbellbacteria
Giovannonibacteria
Wolfebacteria
Jorgensenbacteria
Azambacteria
Parcubacteria
Yanofskybacteria
Moranbacteria
Magasanikbacteria
Uhrbacteria
Falkowbacteria
Candidate
Phyla Radiation
SM2F11
Peregrinibacteria
Gracilibacteria BD1-5, GN02
Absconditabacteria SR1
Saccharibacteria
Berkelbacteria
Woesebacteria
Shapirobacteria
Amesbacteria
Collierbacteria
Pacebacteria
Beckwithbacteria
Roizmanbacteria
Gottesmanbacteria
Levybacteria
Daviesbacteria
Curtissbacteria
Microgenomates
Wirthbacteria
Dojkabacteria WS6
CPR1
CPR3
Katanobacteria
WWE3
Major lineages with isolated representative: italics
Major lineage lacking isolated representative: ●
Micrarchaeota
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Parvarchaeota
DPANN
Pacearchaeota
Nanoarchaeota
Woesearchaeota
Altiarchaeales
Z7ME43
Methanopyri
Methanococci
Hadesarchaea
Thermococci
Methanobacteria
Thermoplasmata
Archaeoglobi
Methanomicrobia
Archaea
Korarch.
Crenarch.
Bathyarc.
YNPFFA
Aigarch.
Halobacteria
Loki.
Thor.
TACK
Thaumarchaeota
Eukaryotes
Opisthokonta
Excavata
Archaeplastida
Chromalveolata
Amoebozoa

Outgroup

- IN an unrooted tree, one cannot tell which is the evolutionary origin
- If you know that a group of sequences is more distant than the rest ('outgroup'), then the root of the tree can be set on the branch separating the outgroup from the rest!


gorilla
orang
chimp
human
gibbon
(outgroup)
Wurzel


Wurzel
gibbon
orang
gorilla
chimp
human