

Simon Haendeler, Ronja Reinhardt 28.01.2022

Evolutionary sequence alignments

How are they made?

What do they tell us?

- What is an evolutionary tree?
- How are multiple sequence alignments made?
- Which parameters are important to consider?

- How old (evolutionarily conserved) is the gene I am working on?
- Which parts of my gene are of core importance for its function?
- Which properties of a region/domain/residue are important for function?
- What is the functional scope of the feature I discovered?
- Which parts of my protein are dependent on each other?

Simon Haendeler

Evolutionary sequence alignments

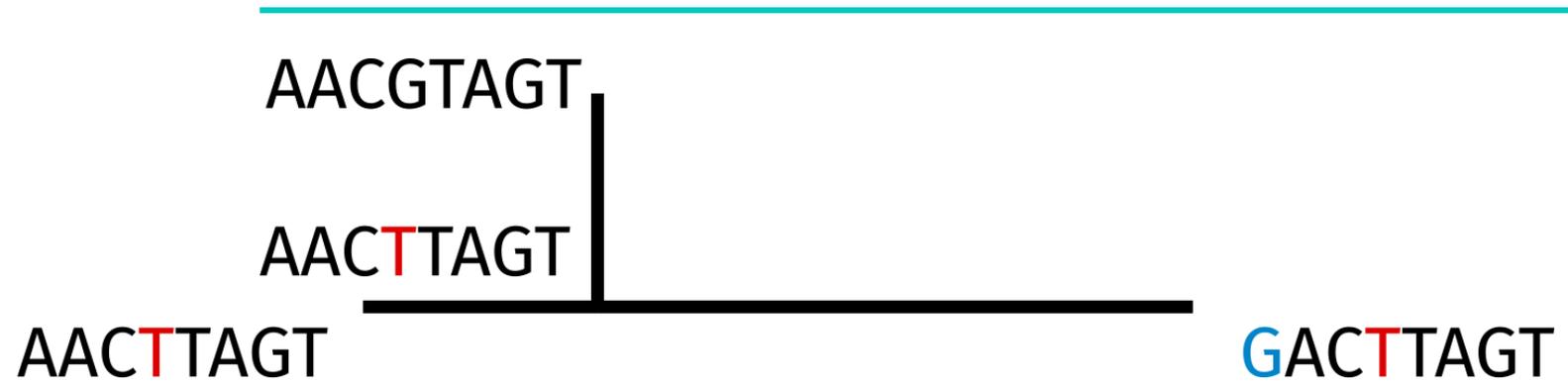
How are they made?

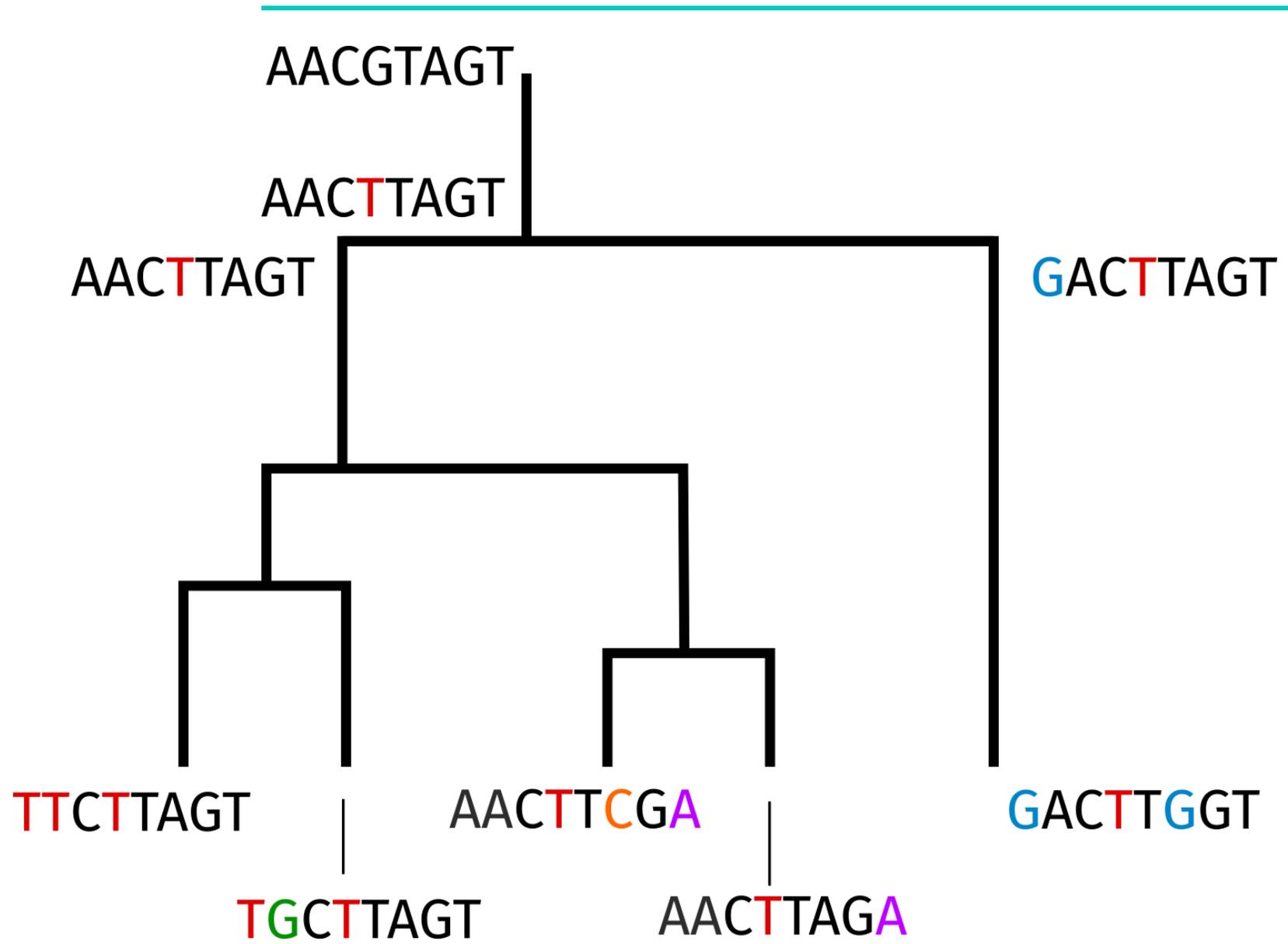
AACGTAGT

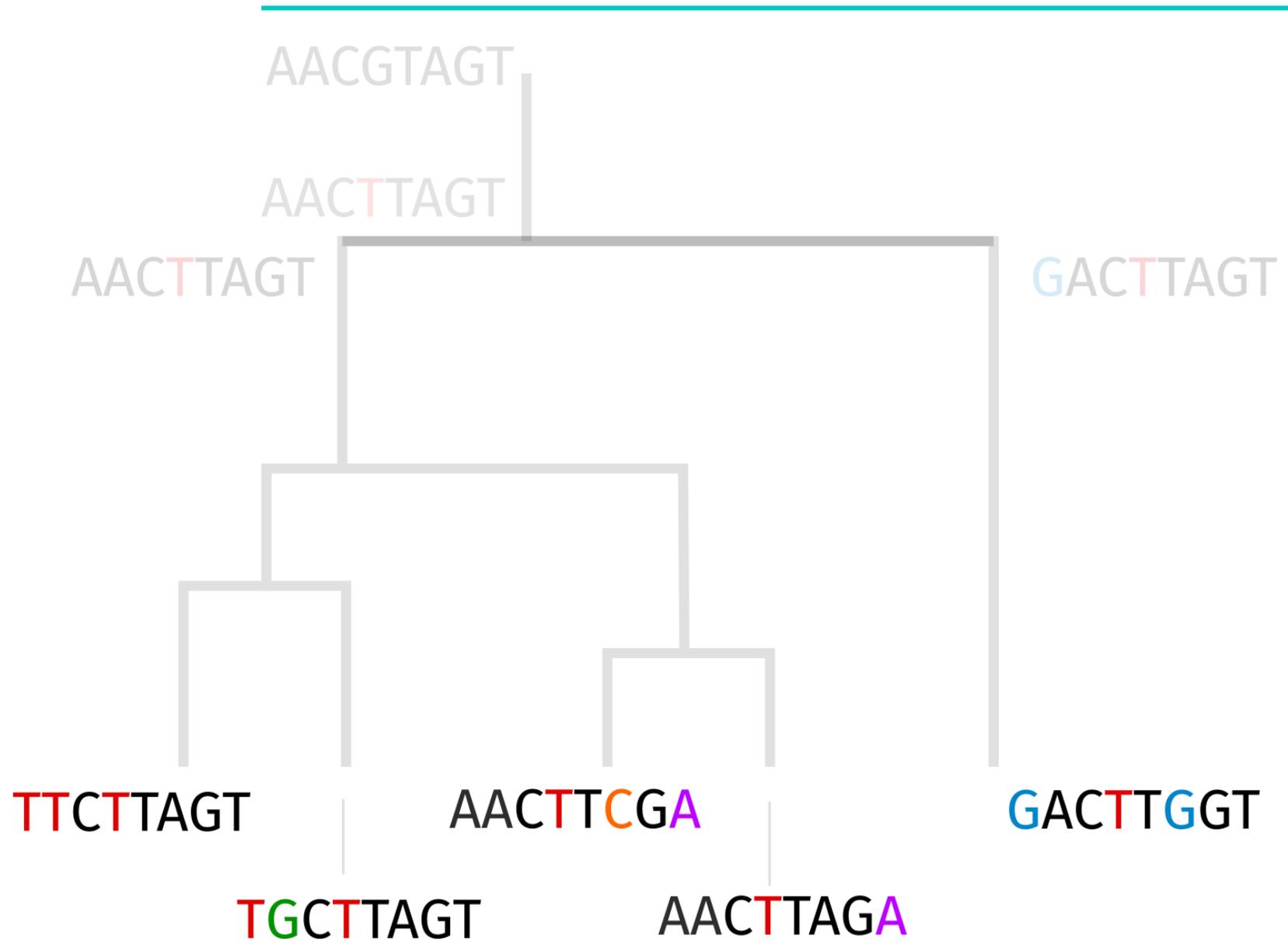
AACGTAGT

AAC**T**TAGT









A

Are you go-ing to Scar - bo-rough Fair? Pars - ley, sage, rose - ma - ry and thyme

B

Are you go-ing to Scar - bo-rough Fair? Pars - ley, sage, rose - ma - ry and thyme

CCGGGE^bF DCGB^b**CB^bCB^bGAFGG**
CCGGGD **E^b**DCGB^b**C** - - **B^bGAF** - **G**

Substitution
 Insertion/deletion
 Regular = Weaker function
 Bold = Stronger function

TTCTTAGT
AACTTCGA

TTCTTAGT
AACTTCGA

TTCTTAGT
AACTTCGA

TTCTTAGT
CTTGT

TTCTTAGT
AACTTCGA

TTCTTAGT
CTTGT

TTCTTAGT
AACTTCGA

TTCTTAGT
CTTGT

TTCTTAGT
CTTGT

TTCTTAGT
AACTTCGA

TTCTTAGT
CTTGT

TTCTTAGT
CTTGT

TTCTTAGT
CTT-GT

TTCTTAGT
AACTTCGA

TTCTTAGT
CT-CGT

TTCTTAGT
CTC-GT

TTCTTAGT
CTTGT

TTCTTAGT
CTTGT

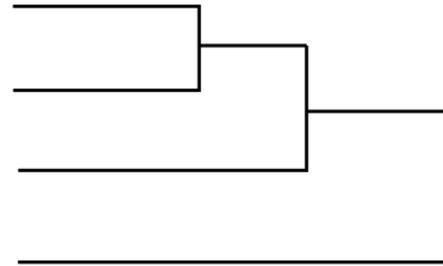
TTCTTAGT
CTT-GT

TTCTTAGT
AATCGA
GACTGT
ACTTAGA
TGCTTAGTAAC

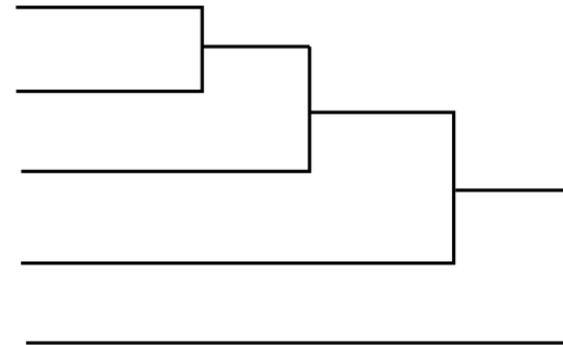
TTCTTAGT
--AATCGA
GACTGT
ACTTAGA
TGCTTAGTAAC



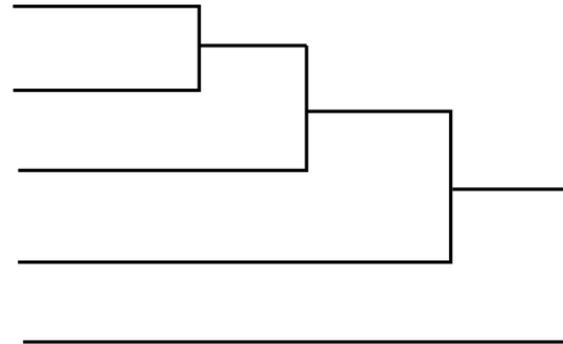
TTCTTAGT
--AATCGA
GAC-T-GT
-ACTTAGA
TGCTTAGTAAC



TTCTTAGT-
--AATCG-A
GAC-T-GT-
-ACTTAG-A
TGCTTAGTAAC



TTCTTAGT-
--AATCGA-
GAC-T-GT-
-ACTTAGA-
TGCTTAGTAAC



BLAST®

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEWS

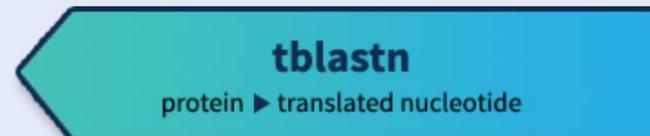
A new feature was added to the NCBI IgBLAST webpage

IgBLAST is now able to determine Ig isotypes

Mon, 01 Nov 2021 12:00:00 EST

[More BLAST news...](#)

Web BLAST



Standard Nucleotide BLAST

blastn

blastp

blastx

tblastn

tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Reset page

Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Query subrange [?](#)

```
>sp|Q62226|SHH_MOUSE Sonic hedgehog protein OS=Mus musculus  
OX=10090 GN=Shh PE=1 SY=2  
M L L L A R C F L V I L A S S L L V C P G L A C G P G R G F G K R R H P K K L T P L A Y K Q F I P N V A E K  
T L G A S
```

From

To

Or, upload file

Browse... No file selected. [?](#)

Job Title

sp|Q62226|SHH_MOUSE Sonic hedgehog protein...

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**New columns added to the
Description Table**

Click 'Select Columns' or 'Manage
Columns'.



Choose Search Set

Database

Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt) [?](#)

Organism

Optional

Enter organism name or id--completions will be suggested exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude

Optional

Models (XM/XP) Uncultured/environmental sample sequences

Limit to

Optional

Sequences from type material

BLAST[®] » **blastp suite** » RID-Z7KN4TA9016

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

Format Request Status

[\[Formatting options\]](#)

Job Title: sp|Q62226|SHH_MOUSE Sonic hedgehog protein

Request ID	Z7KN4TA9016
Status	Searching
Submitted at	Fri Jan 28 06:08:20 2022
Current time	Fri Jan 28 06:08:31 2022
Time since submission	00:00:11

This page will be automatically updated in **2** seconds

- Descriptions
- Graphic Summary
- Alignments
- Taxonomy

Sequences producing significant alignments

Download ▼ New Select columns ▼ Show 100 ▼ ?

select all 100 sequences selected

[GenPept](#)
 [Graphics](#)
 [Distance tree of results](#)
 [Multiple alignment](#)
New [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	sonic hedgehog protein precursor [Mus musculus]	Mus musculus	898	898	100%	0.0	100.00%	437	NP_033196.1
<input checked="" type="checkbox"/>	unnamed protein product [Mus musculus]	Mus musculus	896	896	100%	0.0	99.77%	437	BAC34996.1
<input checked="" type="checkbox"/>	sonic hedgehog protein isoform X1 [Mus caroli]	Mus caroli	892	892	100%	0.0	99.08%	438	XP_021018615.1
<input checked="" type="checkbox"/>	sonic hedgehog protein [Mus pahari]	Mus pahari	885	885	100%	0.0	98.40%	437	XP_021046321.1
<input checked="" type="checkbox"/>	sonic hedgehog protein [Grammomys surdaster]	Grammomys surdaster	876	876	100%	0.0	97.48%	438	XP_028630595.1
<input checked="" type="checkbox"/>	sonic hedgehog protein [Rattus rattus]	Rattus rattus	874	874	100%	0.0	97.48%	437	XP_032763078.1
<input checked="" type="checkbox"/>	sonic hedgehog homolog (Drosophila) [Rattus norvegicus]	Rattus norvegicus	873	873	100%	0.0	97.48%	437	EDL86418.1
<input checked="" type="checkbox"/>	sonic hedgehog [Meriones unguiculatus]	Meriones unguiculatus	867	867	100%	0.0	96.57%	437	BAD30089.1
<input checked="" type="checkbox"/>	sonic hedgehog protein precursor [Rattus norvegicus]	Rattus norvegicus	863	863	100%	0.0	96.57%	437	NP_058917.1
<input checked="" type="checkbox"/>	sonic hedgehog protein [Mastomys coucha]	Mastomys coucha	840	840	100%	0.0	96.57%	437	XP_031236118.1

Clustal Omega

Input form

Web services

Help & Documentation

Bioinformatics Tools FAQ

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile to align **or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

<http://www.ebi.ac.uk/Tools/mseq>

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
NPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSRDRRSKYGM  
LARLAVEAGFDWVYYESKAHIIHCSVKAENSVAAKSGGCFPGSAIVHLEQGGTKLVKDLSPGDRVLAADDQGQLLYSDFLT  
FLDREDGTKKVFYVIETREPRERLLLTAHLLFVAPHNDSVAAWPEPPSSAGARLRSPPGAEGRRALFASRVRPGQRV  
YVVAERDGD RRLLPAAVHSVTLREETTAYAPLTAQGTILINRVLASCYAVIEEHSWAHWAFAPFRLAHALLAALEPSRT  
DRGGGGGSGSNGGRLPSPAPDAAPAPDAAASAAGIHWYSQLLYQIGTWLLDSEALHPLGMAVKSS  
>KAH8209126.1 SHH protein [Marmota monax]  
MLLLARCLLVVLVSSLLVCSGLACGPRGFGKRRHPKLTPLAYKQFIPNVAEKT LGASGRYEGKISRNSERFKELTPNY  
NPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSRDRRSKYGM
```

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

DEALIGN INPUT SEQUENCES

no

MBED-LIKE CLUSTERING GUIDE-TREE

yes

MBED-LIKE CLUSTERING ITERATION

yes

NUMBER of COMBINED ITERATIONS

default(0)

MAX GUIDE TREE ITERATIONS

default

MAX HMM ITERATIONS

default

ORDER

aligned

STEP 3 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Results for job clustalo-l20220128-111810-0664-66042774-p1m

Alignments

Result Summary

Guide Tree

Phylogenetic Tree

Results Viewers

Submission Details

Download Alignment File

Show Colors

CLUSTAL 0(1.2.4) multiple sequence alignment

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 7%

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. [doi: 10.1093/nar/gkw256](https://doi.org/10.1093/nar/gkw256)

Tree Inference

Model Selection

Analysis Results

For a quick start, take a look at the [tutorial](#) for the IQ-TREE web server.

Please visit the [IQ-TREE homepage](#) for more information or if you want to download the main software.

Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file :

C:\fakepath\clustalo-l20220128-1

Browse...

Show example >

Use example alignment:

Yes

?

Sequence type:

Auto-detect

DNA

Protein

Codon

?

DNA->AA

Binary

Morphology

Partition file:

This field is optional.

Browse...

Show example >

Partition type:

Edge-linked

Edge-unlinked

?

Substitution Model Options

Substitution model:

Auto

?

FreeRate heterogeneity:

Yes [+R]

Rate heterogeneity:

Gamma [+G]

Invar. sites [+I]

?

#rate categories:

4

State frequency:

Empirical
(from data)

AA model
(from matrix)

ML-optimized

Codon F1x4

Codon F3x4

**Ascertainment bias
correction:**

Yes [+ASC]

?

IQ-TREE Search Parameters

Perturbation strength:



IQ-TREE stopping rule:



Hide <

Help

Option	Usage and meaning
Perturbation strength [-pers]	Specify perturbation strength (between 0 and 1) for randomized nearest neighbor interchange (NNI). <i>DEFAULT: 0.5</i>
Stopping rule [-numstop]	Specify number of unsuccessful (≥ 100) iterations to stop. <i>DEFAULT: 100</i>

NOTICE: While the default parameters were empirically determined to work well under our extensive benchmark ([Nguyen et al., 2015](#)), it might not hold true for all data sets. If in doubt that tree search is still stuck in local optima, one should repeat analysis with at least 10 IQ-TREE runs. Moreover, our experience showed that -pers and -numstop are the most relevant options to change in such case. For example, data sets with many short sequences should be analyzed with smaller perturbation strength (-pers) and larger -numstop.

Tree Inference Model Selection **Analysis Results**

User name or Email: QUERY STATUS

<input checked="" type="checkbox"/>	No.	Submission Time	Status
<input checked="" type="checkbox"/>	1	2022-01-28 12:22	Success

Summary Run Log Full Result

Please bookmark the following link to later monitor/retrieve results:
<http://iqtree.cibiv.univie.ac.at/?user=guest&jobid=220128122218>

If you want to monitor the progress, click on Run Log above. If you hit QUERY STATUS, the page is reloaded.
You can [download IQ-TREE](#) and run it locally with the command-line:

```
path_to_iqtree -s clustalo-I20220128-111810-0664-66042774-p1m.clustal_num -m TEST -bb 1000 -alrt 1000
```

Note: The CPU time limit is 24 hours and RAM limit is 1GB. Your job will be stopped if it exceeds these limits.
(In that case, please download the stopped job and use the above command-line to resume the run from the last checkpoint on your local PC as described [here](#).)

MAXIMUM LIKELIHOOD TREE

Log-likelihood of the tree: -8886.1386 (s.e. 374.6907)
Unconstrained log-likelihood (without tree): -5022.4931
Number of free parameters (#branches + #model parameters): 218
Akaike information criterion (AIC) score: 18208.2772
Corrected Akaike information criterion (AICc) score: 18350.1554
Bayesian information criterion (BIC) score: 19253.2528

Total tree length (sum of branch lengths): 9.6887
Sum of internal branch lengths: 4.3598 (44.9986% of tree length)

Tree in newick format:

```
(NP_001185482.1:0.1078522072, ((XP_031795387.1:0.0890900353, (XP_02769
```

CONSENSUS TREE

Consensus tree is constructed from 1000bootstrap trees

Log-likelihood of consensus tree: -8887.928703

Robinson-Foulds distance between ML tree and consensus tree: 6

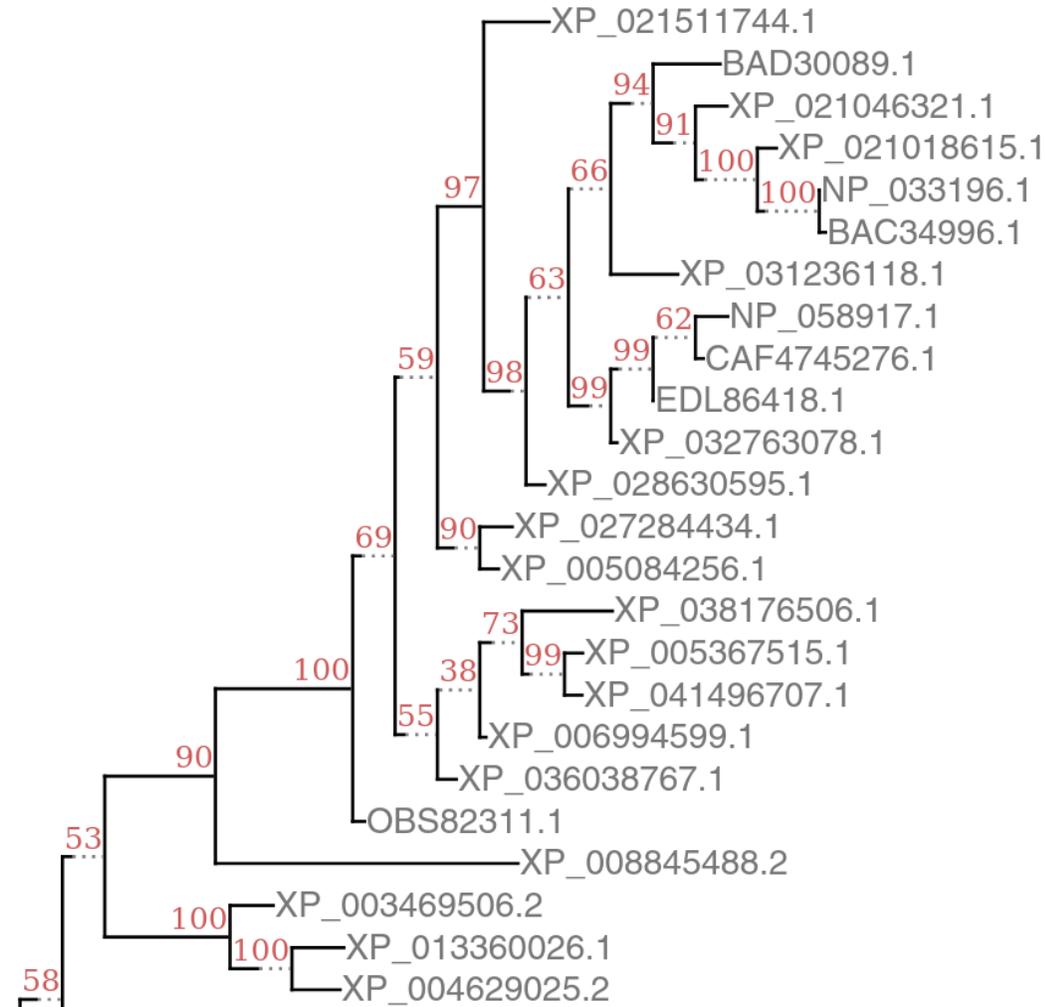
Branches with support >0.000000% are kept (extended consensus)

Branch lengths are optimized by maximum likelihood on original alignment

Numbers in parentheses are bootstrap supports (%)

Consensus tree in newick format:

```
(NP_001185482.1:0.1086401941, ((XP_031795387.1:0.089726407
```



Ronja Reinhardt

Evolutionary sequence alignments

What do they tell us?

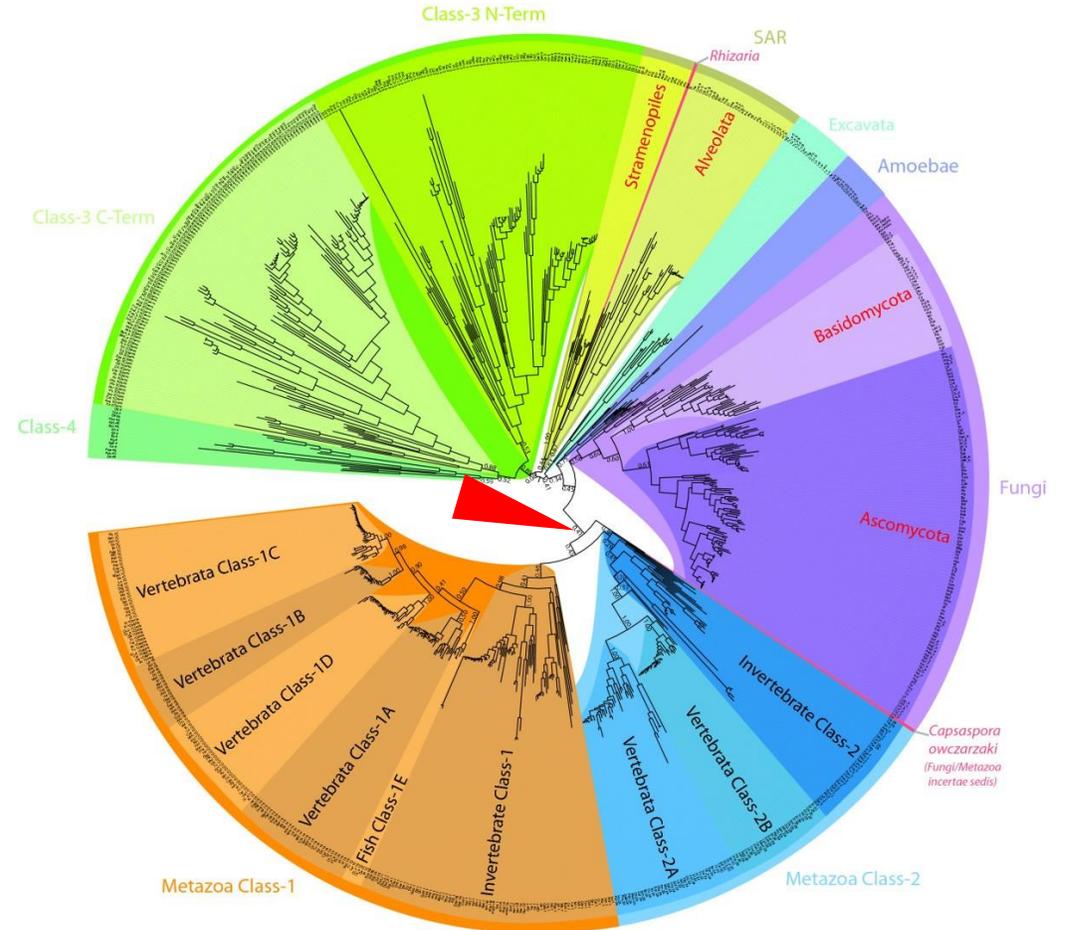
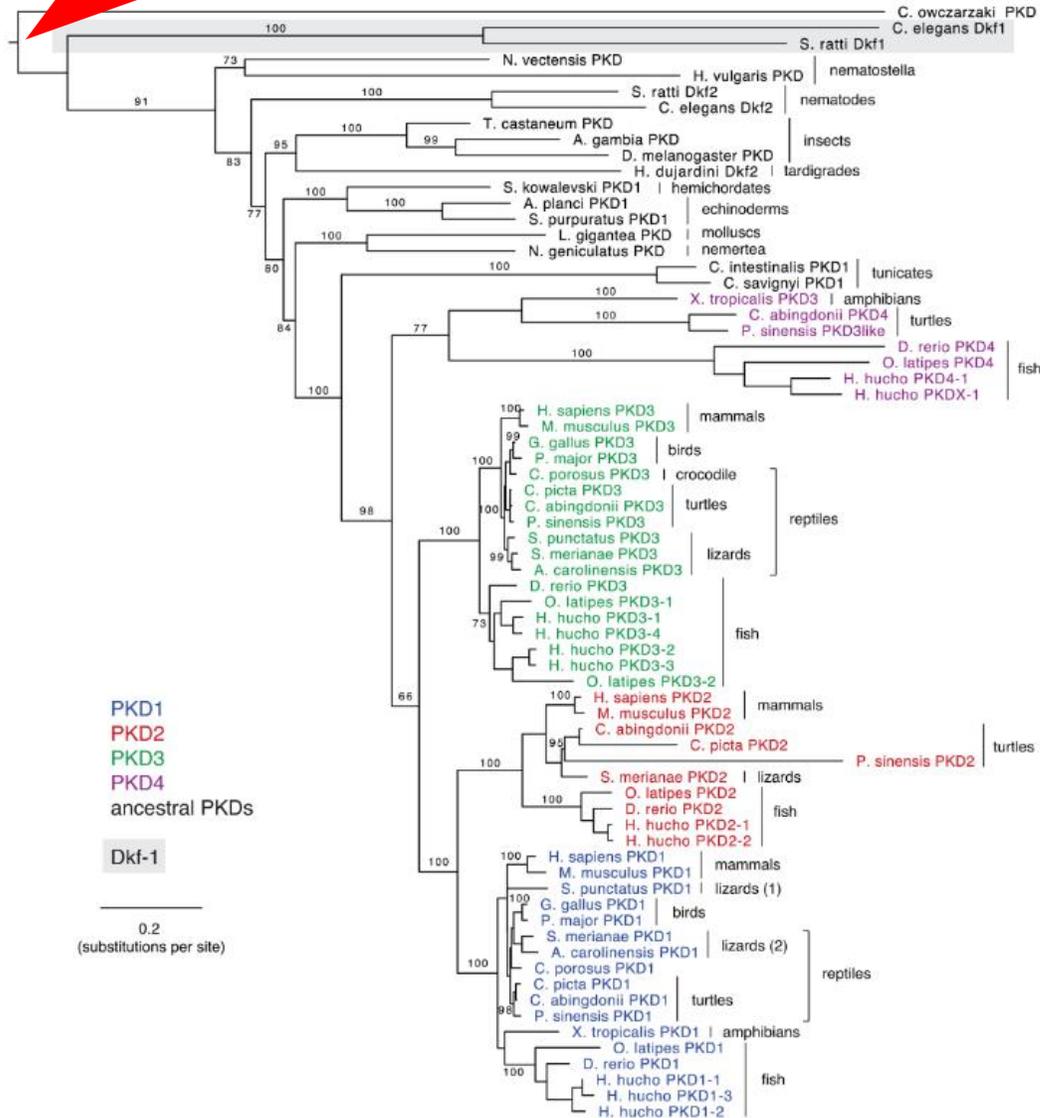
- How old (evolutionarily conserved) is the gene I am working on?
- Which parts of my gene are of core importance for its function?
- Which properties of a region/domain/residue are important for function?
- What is the functional scope of the feature I discovered?
- Which parts of my protein are dependent on each other?

– Why am I curious about it?

➤ What is the functional scope of my discovery?

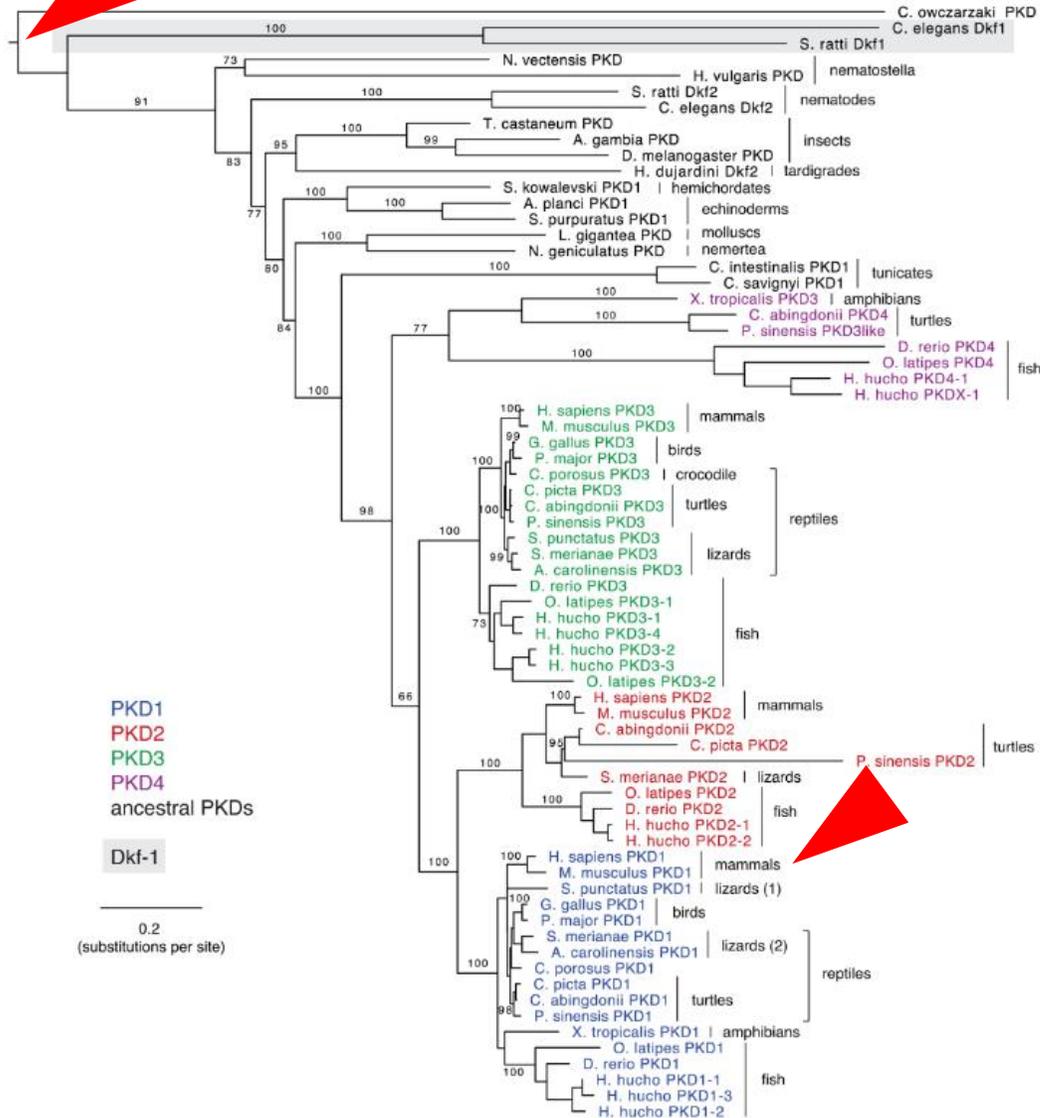
- Is a gene conserved in every living cell?
- Is my feature only conserved in human?
- Or maybe only in my mouse cell line?
- Is there a compensation for it in other organisms?

How old is the gene I am working on?



This protein/gene is conserved in **all animals!**

How old is the gene I am working on?



TIMETREE
THE TIMESCALE of LIFE

Search Pair Time

Taxon 1:

Taxon 2:

Clear Search

Capsaspora owczarzaki

Capsaspora owczarzaki

Versus

human

Homo sapiens

Median Time:
928 MYA

This protein/gene was conserved over **900 mio years of evolution!**

Which parts are of core importance?

– Why am I curious about this?

- Discovery of domains/motifs
- Construct design

– How do I find out?

- Look at conserved parts in an alignment
- Check on Uniprot

Which parts are of core importance?

– How do I find out?

Jalview 2.11.1.7

File Tools Help Window

C:\Users\Ronja\Desktop\Share backup 20200225\PHD\Sequences\Alignments\evolutionar...

File Edit Select View Annotations Format Colour Calculate Web Servi

- New View Ctrl+T 600
- Expand Views X
- Gather Views G
- Show >
- Hide >
- Automatic Scrolling
- Show Sequence Features
- Feature Settings...
- Sequence ID Tooltip >
- Alignment Properties
- Overview Window**

CoPKD1/1-1157
CeDkf1/1-722
SrDkf1/1-813
PmPKD1/1-837
HvPKD1/1-737
GiPKD1/1-793
AsPKD1/1-770
TcPKD1/1-805
DmPKD1/1-836
SlkPKD1/1-859
DfrPKD1/1-878
XtPKD1/1-860
GgPKD1/1-891
HsPKD1/1-912
MmPKD1/1-918

Conservati
Quality
Consensus
Occupancy

Consensus: PSESF IGR++RSNSQSY IGRP I WTHSSLDK I L L S S + KVKVPHTF V I HSYTRPTVCQYCKLL KGLFRQGLQCKDCKFN

Overview C:\Users\Ronja\Desktop\Share backup 20200225\PHD\Sequences\Alignments\evolutionar...

1 2 3 4 5

This protein has 5 domains.

Which parts are of core importance?

– How do I find out?

UniProtKB - Q15139 (KPCD1_HUMAN)

Display [Help video](#) [BLAST](#) [Align](#) [Format](#) [Added to basket](#) [History](#)

Entry Protein **Serine/threonine-protein kinase D1**

Publications [Gene](#) **PRKD1**

Family & Domainsⁱ

Domains and Repeats

Feature key	Position(s)	Description
Domain ⁱ	422 – 541	PH PROSITE-ProRule annotation
Domain ⁱ	583 – 839	Protein kinase PROSITE-ProRule annotation

Region

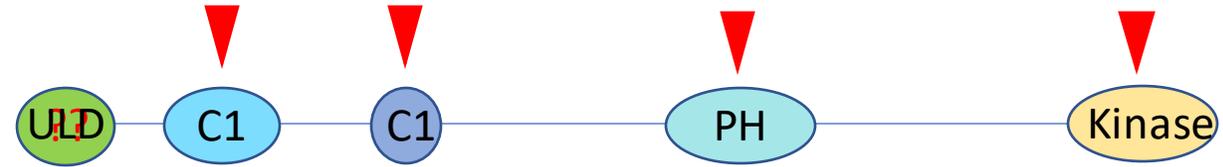
Feature key	Position(s)	Description
Region ⁱ	377 – 402	Disordered Sequence analysis

Sequence similaritiesⁱ

Belongs to the protein kinase superfamily, CAMK Ser/Thr protein kinase family, PKD subfamily. [Curated](#)

Zinc finger

Feature key	Position(s)	Description
Zinc finger ⁱ	146 – 196	Phorbol-ester/DAG-type 1 PROSITE-ProRule annotation
Zinc finger ⁱ	270 – 320	Phorbol-ester/DAG-type 2 PROSITE-ProRule annotation



This protein has 5 domains.

UniProt reports only 4.

New domain discovered ✓

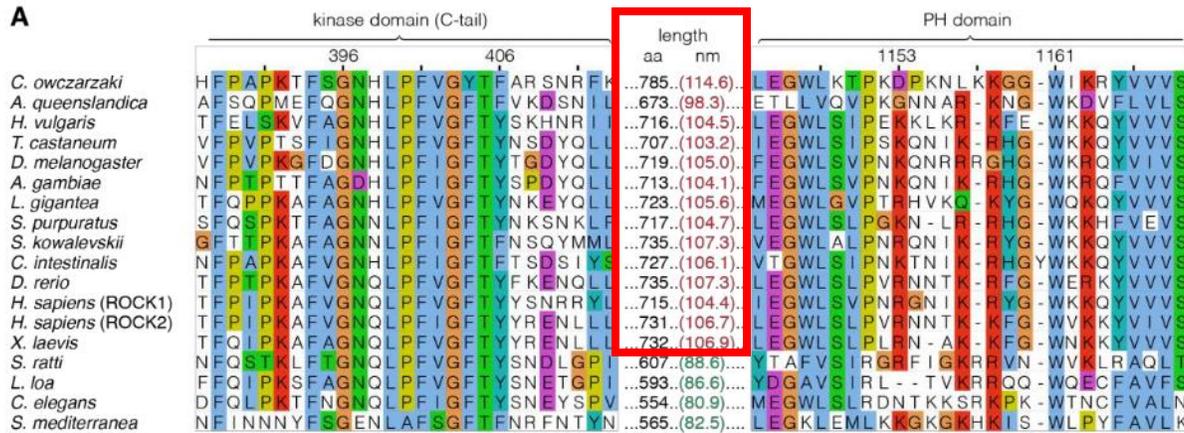
Elsner et al. 2019

– Why am I curious about this?

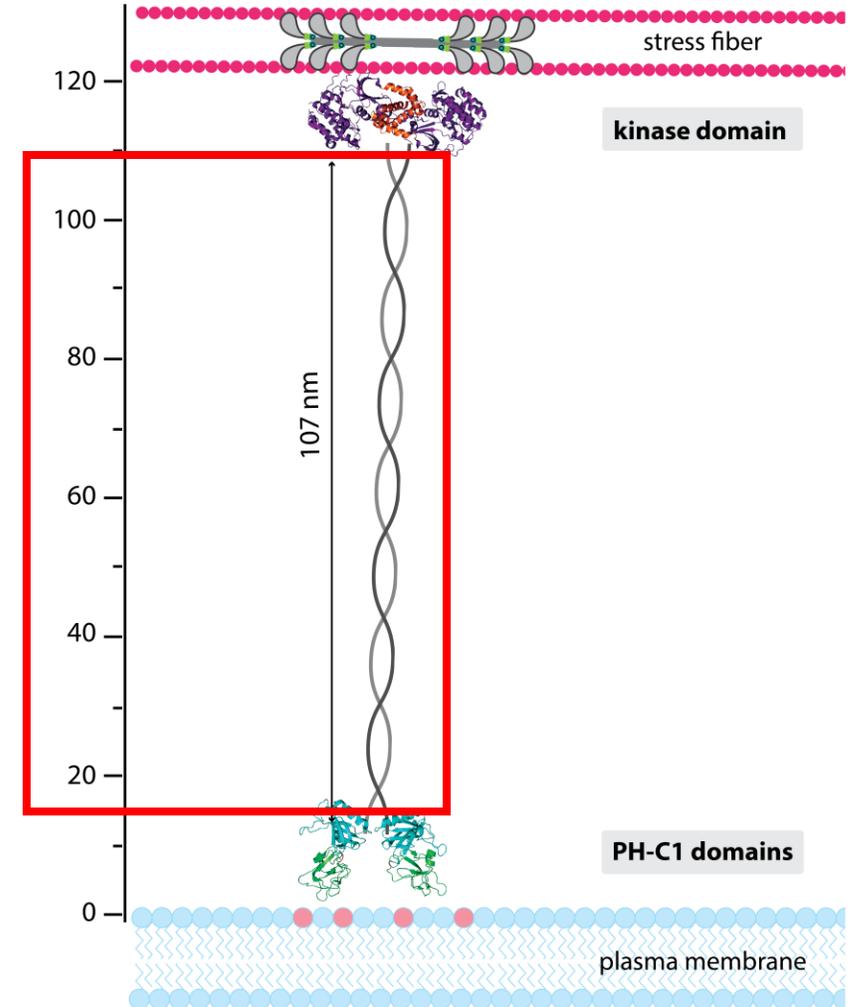
A conserved property is always an important property!

- Length
- Charge
- Hydrophobicity
- ...

- Conserved length



This protein can only function if the catalytic domain is held in the right distance from the membrane (*Trübstein et al. 2015*).



Which properties are important?

- Charge

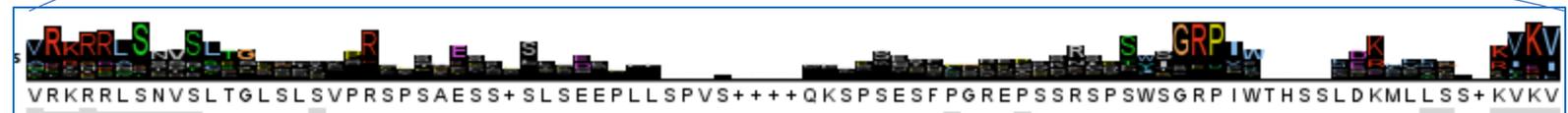


I don't know why this is important but I am sure it is!!

pI: 3.44 – 4.36
length: 14-100



pI: 9.69 -11.45
Length: 35-73

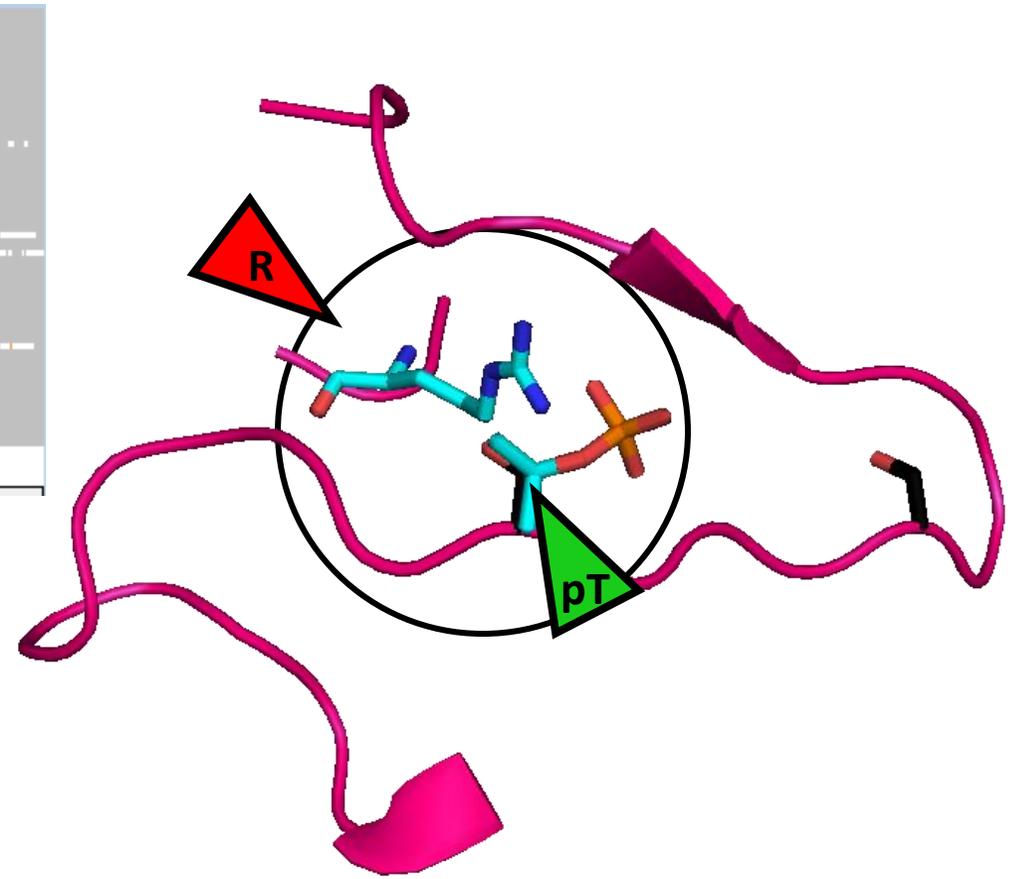
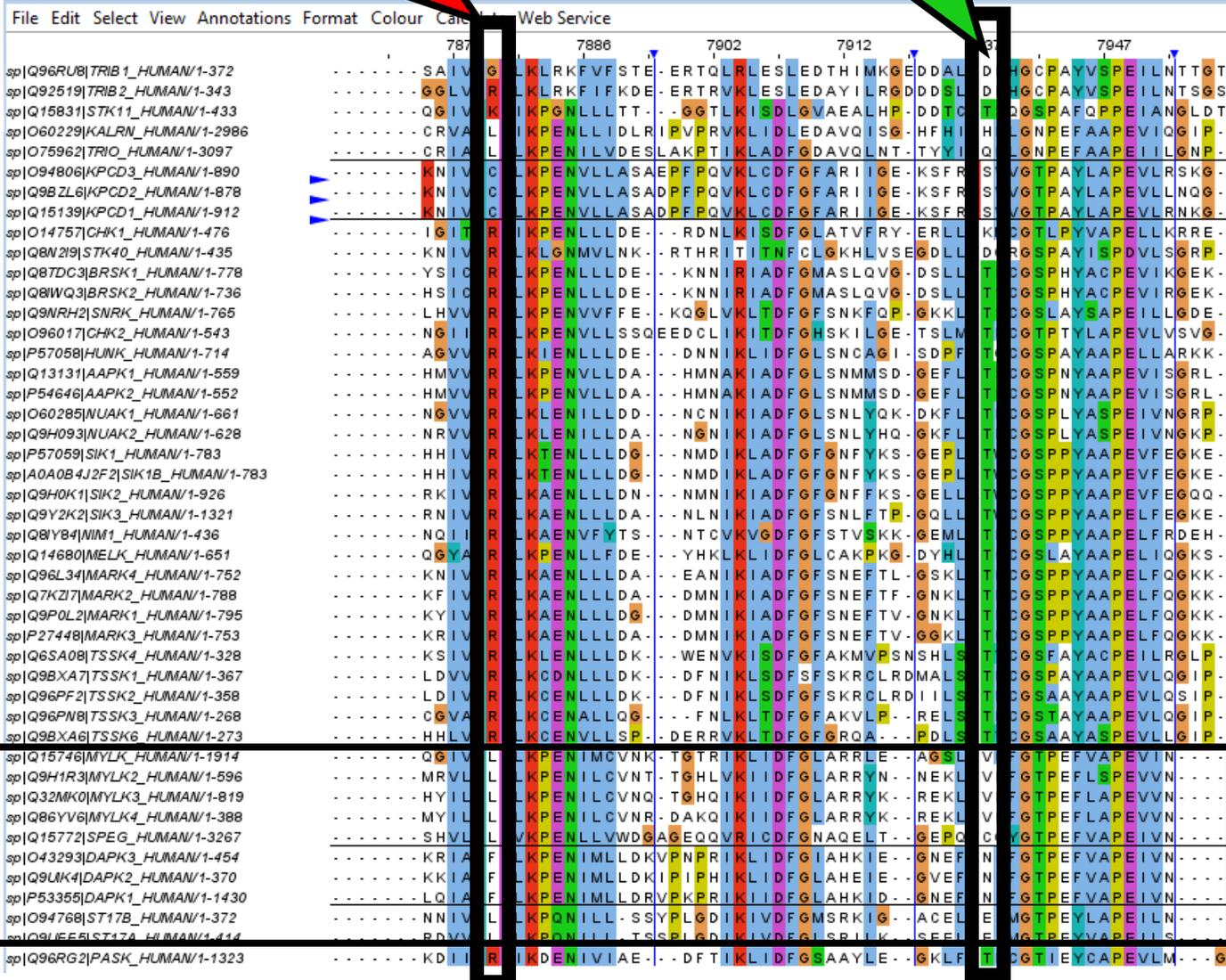


Which parts are dependent on each other?

– Why am I curious about this?

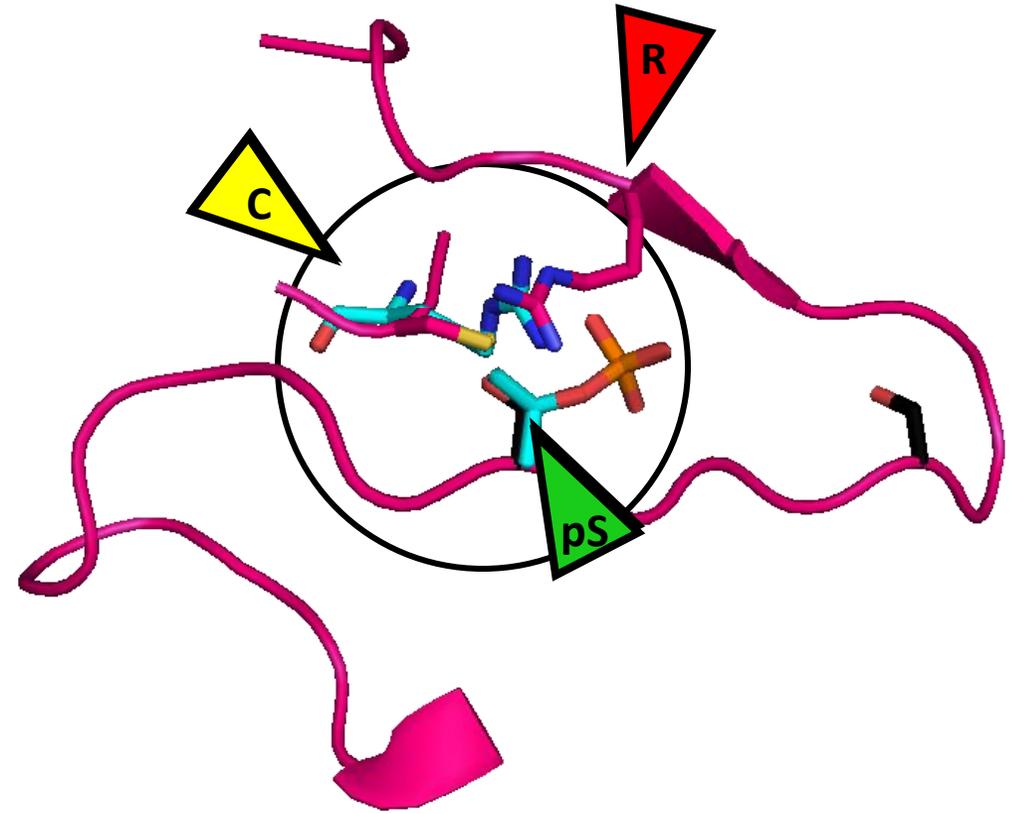
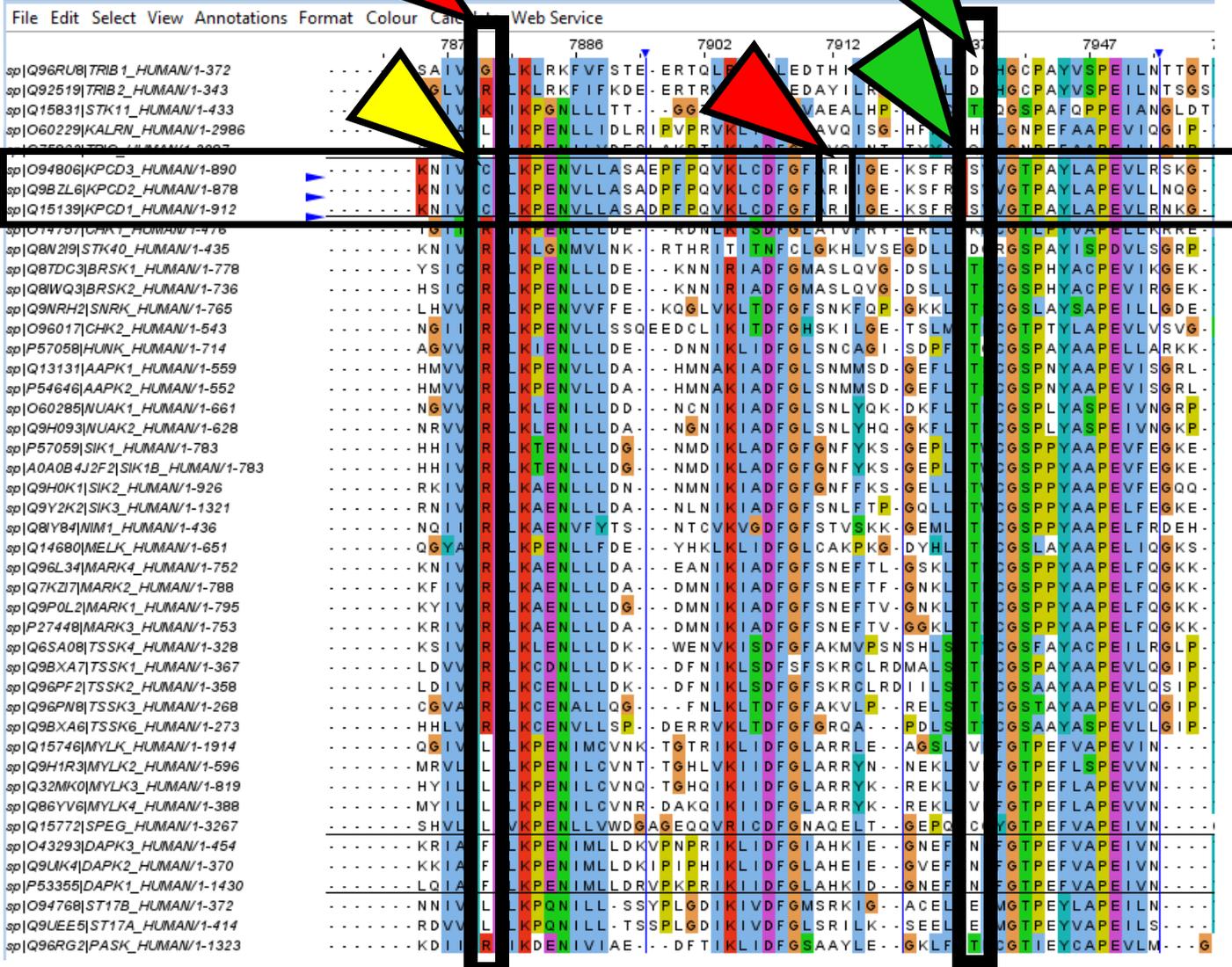
- Functionally connected residues evolve together
- Prediction and verification of intramolecular interactions

Which parts are dependent on each other?



Functionally connected residues evolve together

Which parts are dependent on each other?



Important features don't get lost in evolution!

- They can visualize the traces of evolutionary selection!

Useful to identify:

- phylogenetic relationships
- function/ structure based on similarity
- function based on conserved properties
- functional connection based on coevolution
- relevance of discovered features

Thank you for your attention!

If you ever think our experience can be of use for you.

Find us and ask 😊

Multiple sequence alignment: ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)

Coevolution analysis: CAPS (<http://caps.tcd.ie/caps/analysis.html>)

Phylogenetic trees: IQ-Tree (<http://www.iqtree.org/>)

Evolutionary time: Timetree (<http://www.timetree.org/>)

Interactive alignment multitool: ProViz (<http://slim.icr.ac.uk/proviz/index.php>)

Tools