MAX PERUTZ LABS VIENNA

Simon Haendeler, Ronja Reinhardt 28.01.2022

Evolutionary sequence alignments

How are they made?

What do they tell us?

MAX PERUTZ LABS VIENNA

How are they made?

– What is an evolutionary tree?

– How are multiple sequence alignments made?

– Which parameters are important to consider?

MAX PERUTZ LABS VIENNA

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?

MAX PERUTZ LABS VIENNA

Simon Haendeler

Evolutionary sequence alignments

How are they made?

MAX PERUTZ LABS VIENNA

Evolutionary Trees

AACGTAGT

MAX PERUTZ LABS VIENNA

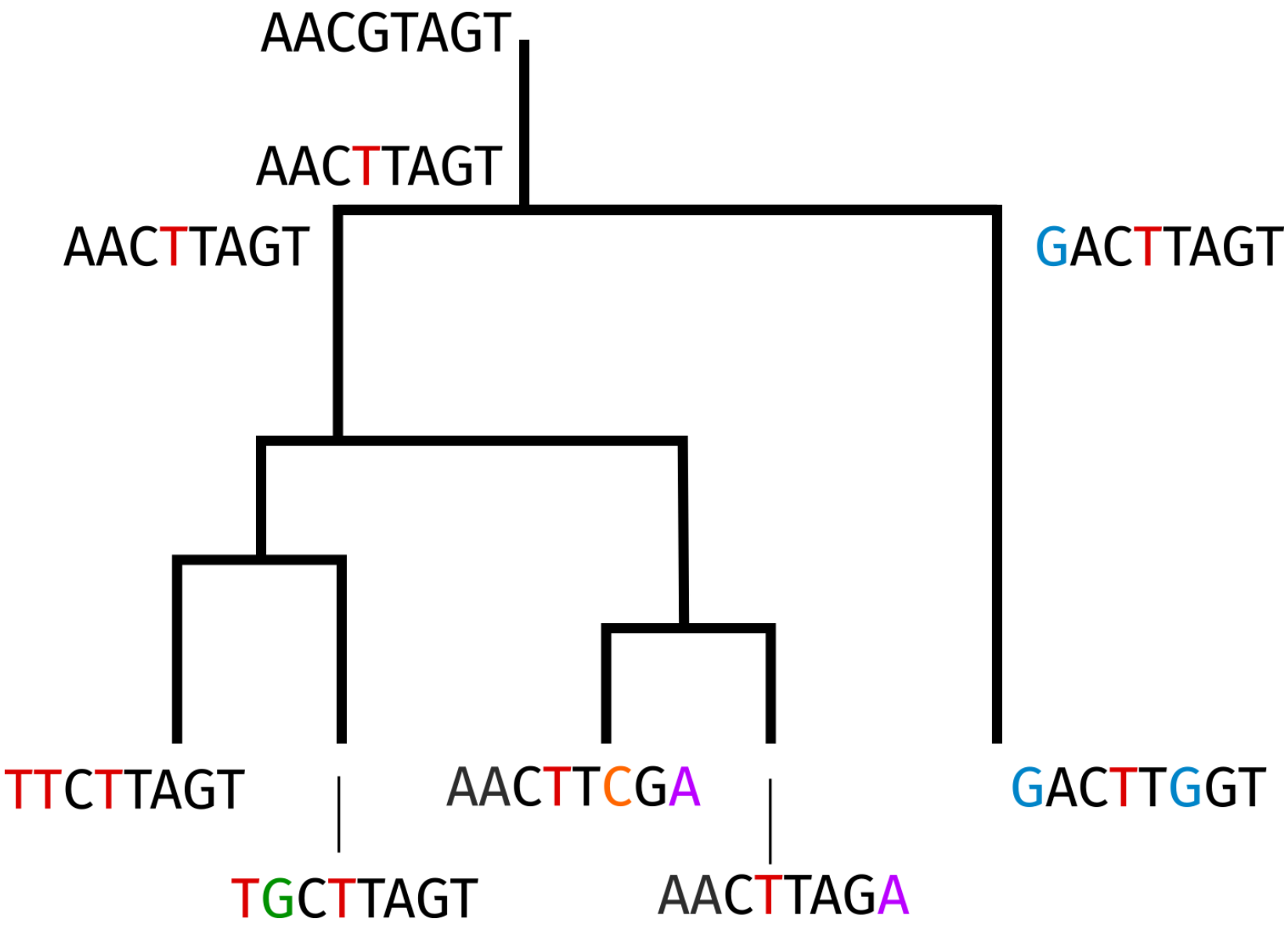
Evolutionary Trees

MAX PERUTZ LABS VIENNA

Evolutionary Trees

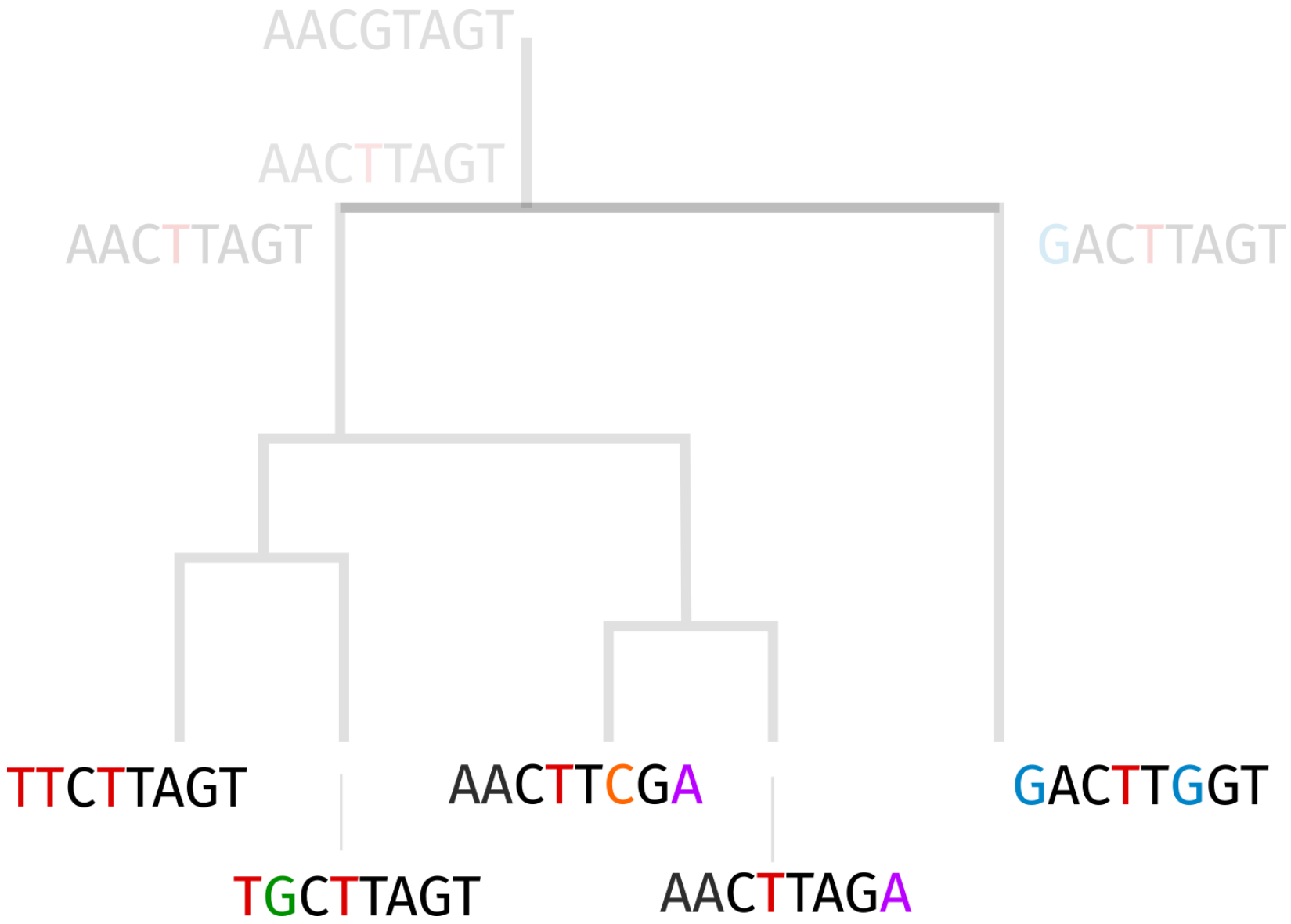
MAX PERUTZ LABS VIENNA

Evolutionary Trees



MAX PERUTZ LABS VIENNA

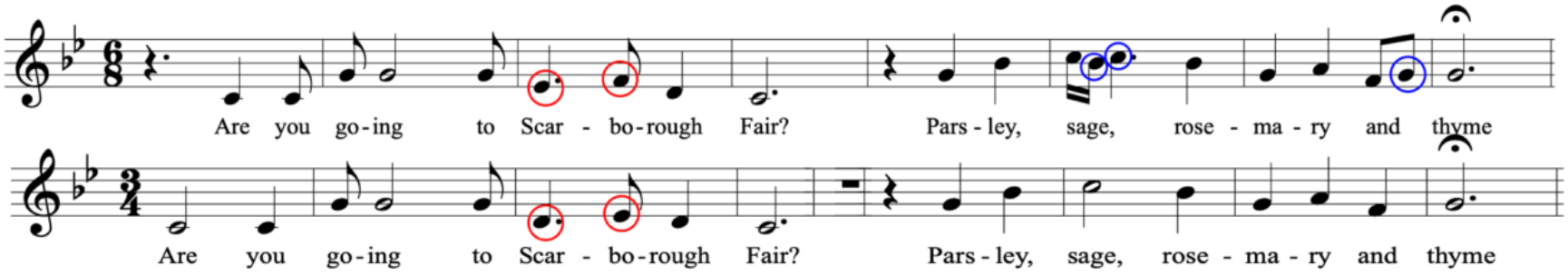
Evolutionary Trees



MAX PERUTZ LABS VIENNA

Evolutionary Trees

A



B

C C G G G E♭ F D C G B♭ C B♭ C B♭ G A F G G

C C G G G D E♭ D C G B♭ C – – B♭ G A F – G

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

MAX PERUTZ LABS VIENNA

Alignments

TTCTTAGT
AACTTCGA

MAX PERUTZ LABS VIENNA

Alignments

TTCTTAGT
AACTTCGA

MAX PERUTZ LABS VIENNA

Alignments

```
TTCTTAGT
AACTTCGA
```

```
TTCTTAGT
CTTGT
```

MAX PERUTZ LABS VIENNA

Alignments

```
TTCTTAGT
AACTTCGA
```

```
TTCTTAGT
CTTGT
```

MAX PERUTZ LABS VIENNA

Alignments

```
TTCTTAGT
AACTTCGA
```

```
TTCTTAGT
CTTGT
```

```
TTCTTAGT
  CTTGT
```

MAX PERUTZ LABS VIENNA

Alignments

```
TTCTTAGT
AACTTCGA
```

```
TTCTTAGT
CTTGT
```

```
TTCTTAGT
  CTTGT
```

```
TTCTTAGT
  CTT-GT
```

MAX PERUTZ LABS VIENNA

Alignments

TTCTTAGT
AACTTCGA

TTCTTAGT
CTTGT

TTCTTAGT
CT-CGT

TTCTTAGT
CTC-GT

TTCTTAGT
CTTGT

TTCTTAGT
CTT-GT

MAX PERUTZ LABS VIENNA

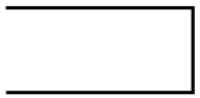
Multiple Sequence Alignments

```
TTCTTAGT
AATCGA
GACTGT
ACTTAGA
TGCTTAGTAAC
```

MAX PERUTZ LABS VIENNA

Multiple Sequence Alignments

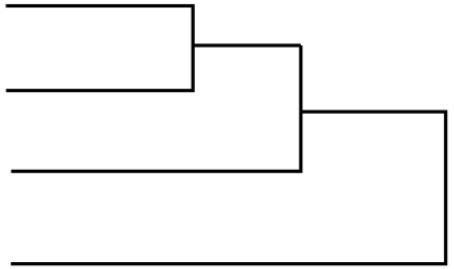
```
TTCTTAGT
--AATCGA
GACTGT
ACTTAGA
TGCTTAGTAAC
```



MAX PERUTZ LABS VIENNA

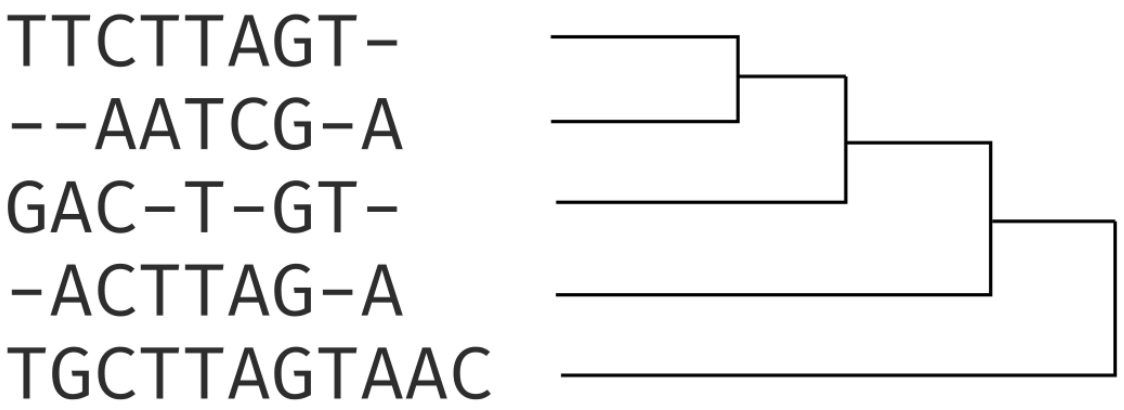
Multiple Sequence Alignments

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



MAX PERUTZ LABS VIENNA

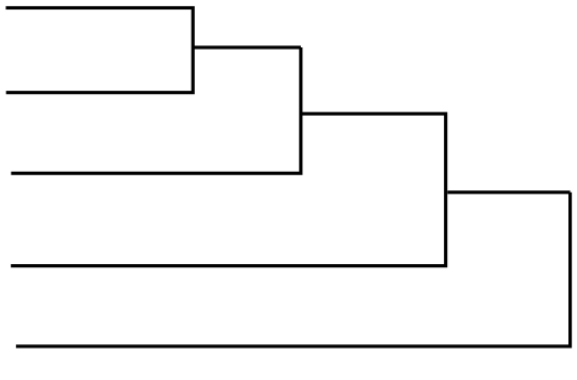
Multiple Sequence Alignments



MAX PERUTZ LABS VIENNA

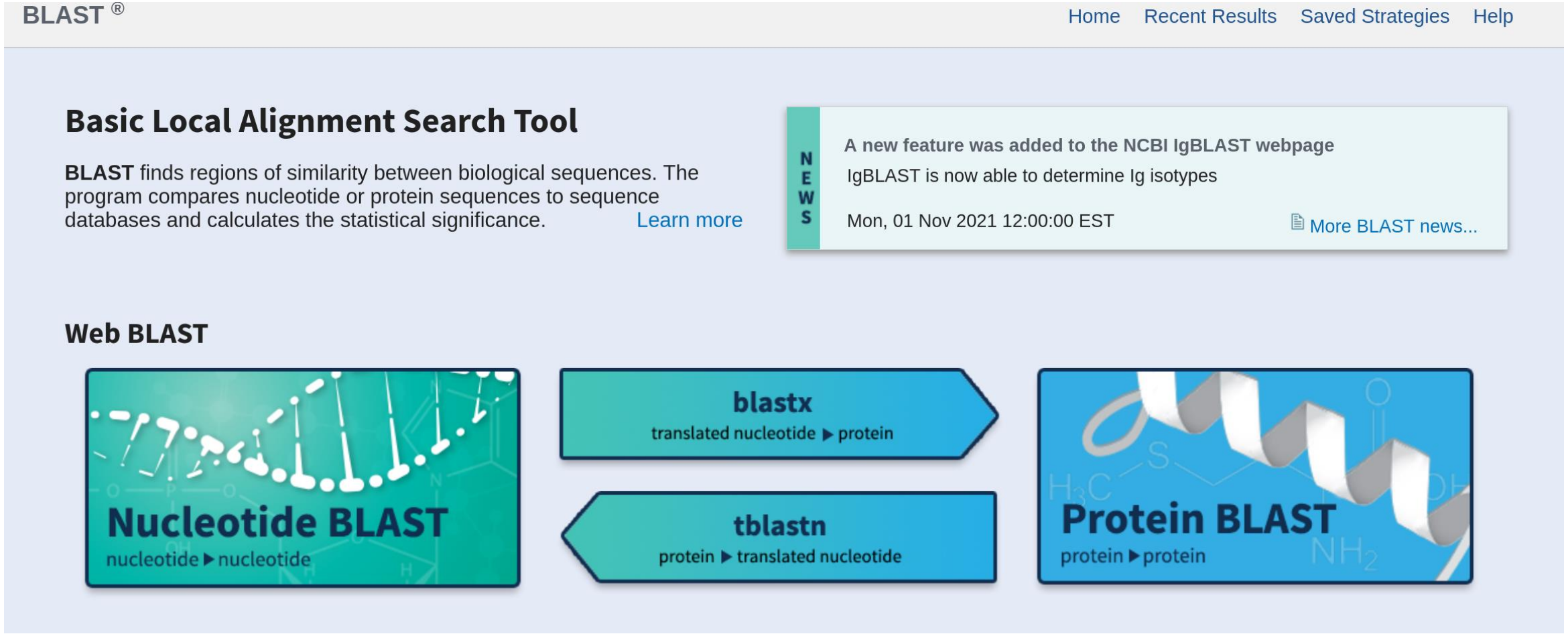
Multiple Sequence Alignments

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



MAX PERUTZ LABS VIENNA

Getting Your Sequences



MAX PERUTZ LABS VIENNA

Getting Your Sequences

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

```
>sp|Q62226|SHH_MOUSE Sonic hedgehog protein OS=Mus musculus
OX=10090 GN=Shh PE=1 SV=2
MLLLLARCFLVILASSLLVCPGLACGPGRGFGKRRHPKKLTPLAYKQFIPNVAEK
TLGAS
```

Query subrange

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

MAX PERUTZ LABS VIENNA

Getting Your Sequences

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

MAX PERUTZ LABS VIENNA

Getting Your Sequences

Descriptions | Graphic Summary | Alignments | Taxonomy

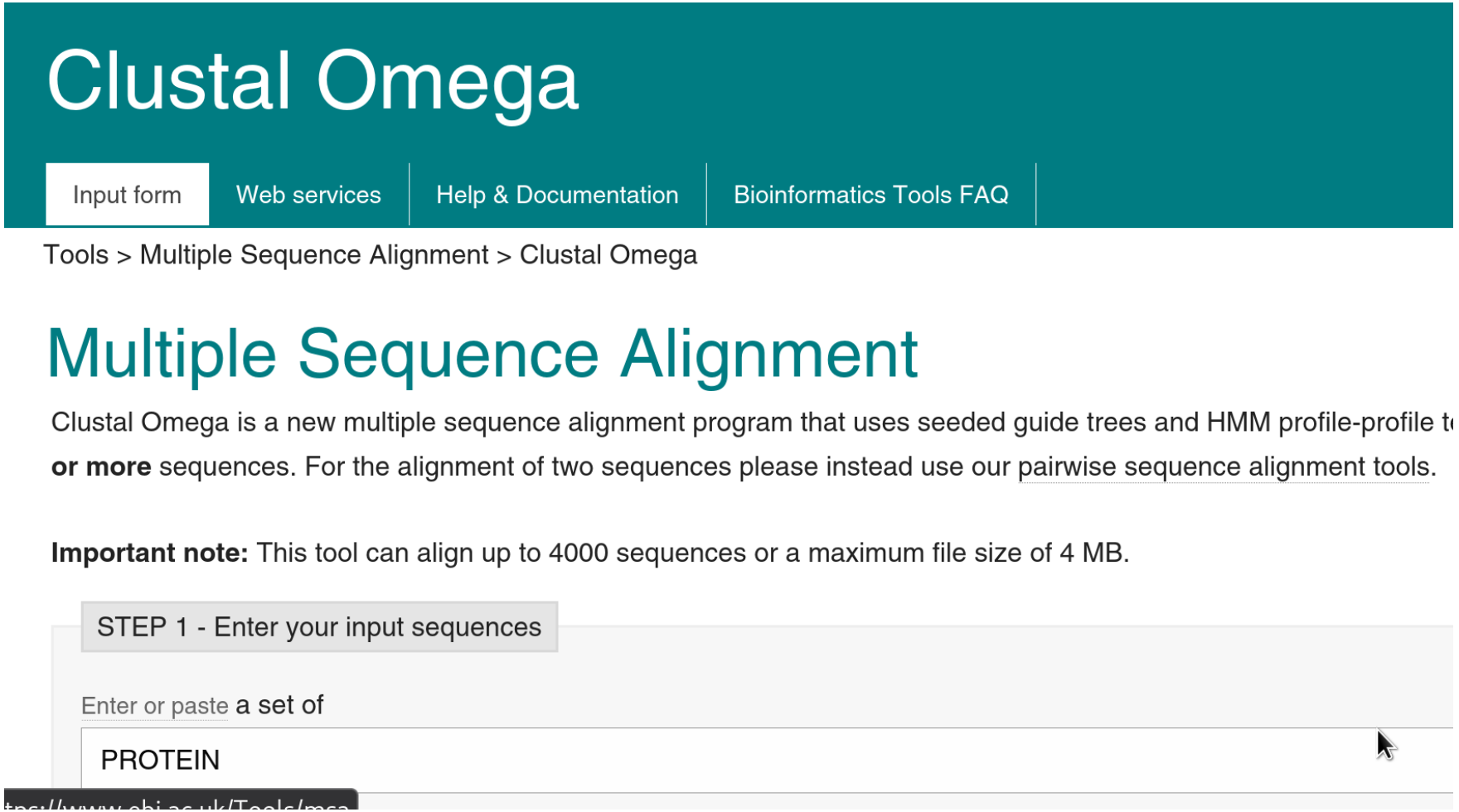
Sequences producing significant alignments Download | New Select columns | Show 100

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
sonic hedgehog protein precursor [Mus musculus]	Mus musculus	898	898	100%	0.0	100.00%	437	NP_033196.1
unnamed protein product [Mus musculus]	Mus musculus	896	896	100%	0.0	99.77%	437	BAC34996.1
sonic hedgehog protein isoform X1 [Mus caroli]	Mus caroli	892	892	100%	0.0	99.08%	438	XP_021018615.1
sonic hedgehog protein [Mus pahari]	Mus pahari	885	885	100%	0.0	98.40%	437	XP_021046321.1
sonic hedgehog protein [Grammomys surdaster]	Grammomys surdaster	876	876	100%	0.0	97.48%	438	XP_028630595.1
sonic hedgehog protein [Rattus rattus]	Rattus rattus	874	874	100%	0.0	97.48%	437	XP_032763078.1
sonic hedgehog homolog (Drosophila) [Rattus norvegicus]	Rattus norvegicus	873	873	100%	0.0	97.48%	437	EDL86418.1
sonic hedgehog [Meriones unguiculatus]	Meriones unguiculatus	867	867	100%	0.0	96.57%	437	BAD30089.1
sonic hedgehog protein precursor [Rattus norvegicus]	Rattus norvegicus	863	863	100%	0.0	96.57%	437	NP_058917.1
sonic hedgehog protein [Mastomys coucha]	Mastomys coucha	840	840	100%	0.0	96.57%	437	XP_031236118.1

MAX PERUTZ LABS VIENNA

Aligning Your Sequences



MAX PERUTZ LABS VIENNA

Aligning Your Sequences

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
NPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRSKYGM
LARLAVEAGFDWVYYESKAHIHCSVKAENSVAAKSGGCFPGSAIVHLEQGGTKLVKDLSPGDRVLAADDQGQLLYSDFLT
FLDREDGTKKVFYVIETREPRERLLLTAAHLLFVAPHNDSVAAWEPEPPSSAGARLRSPPGGAEGRRALFASRVRPGQRV
YVVAERDGDRRLLPAAVHSVTLREETTGAYAPLTAQGTILINRVLASCYAVIEEHSWAHWAFAPFRLAHALLAALEPSRT
DRGGGGGSGSNGGGRLPSPAPDAAPAPDAAASAAGIHWYSQLLYQIGTWLLDSEALHPLGMAVKSS
>KAH8209126.1 SHH protein [Marmota monax]
MLLLARCLLVVLVSSLLVCSGLACGPGRGFGKRRHPKKLTPLAYKQFIPNVAEKTLGASGRYEGKISRNSERFKELTPNY
NPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPGVKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRSKYGM
```

MAX PERUTZ LABS VIENNA

Aligning Your Sequences

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUMBER of COMBINED ITERATIONS
no	yes	yes	default(0)

MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER
default	default	aligned

MAX PERUTZ LABS VIENNA

Aligning Your Sequences

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

MAX PERUTZ LABS VIENNA

Aligning Your Sequences

Results for job clustalo-I20220128-111810-0664-66042774-p1m

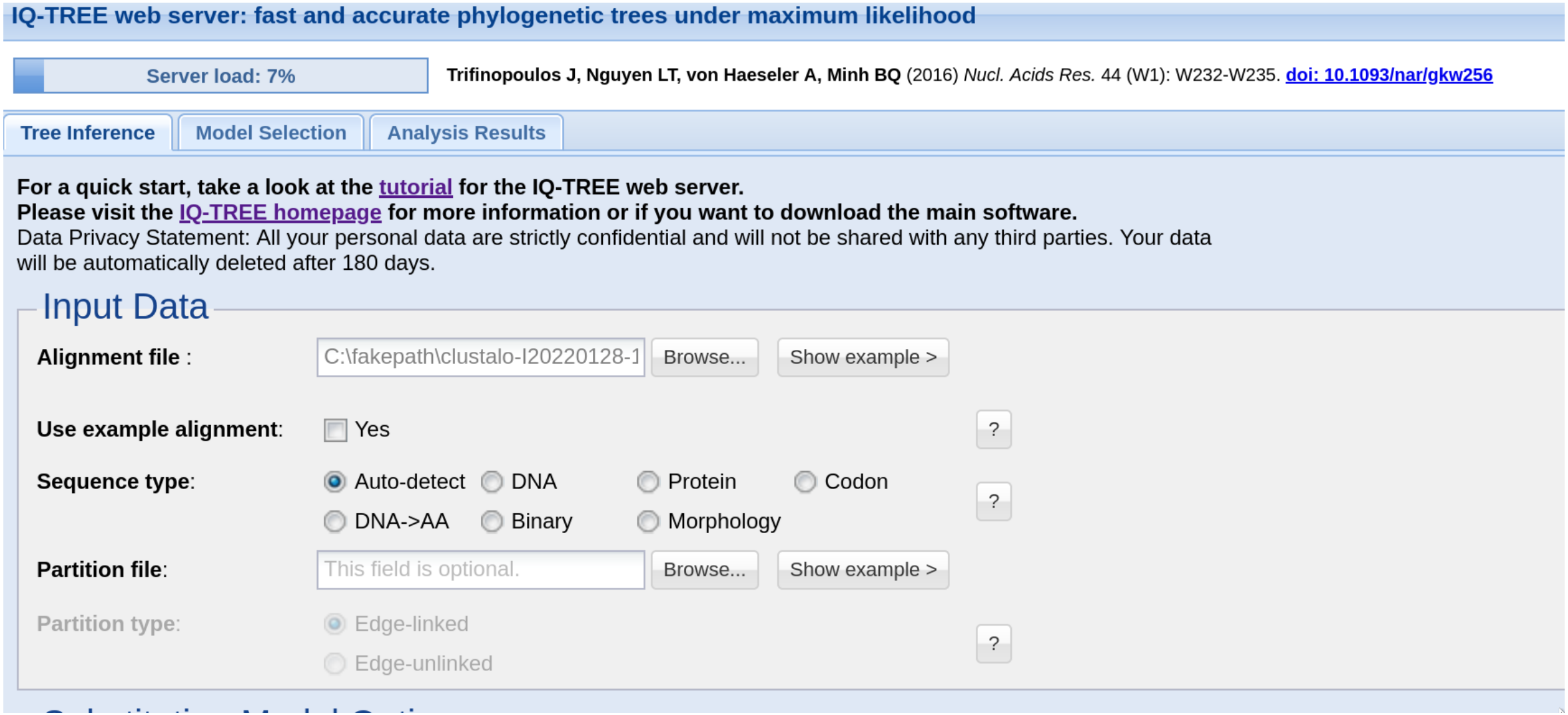
Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Alignment File | Show Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment
```

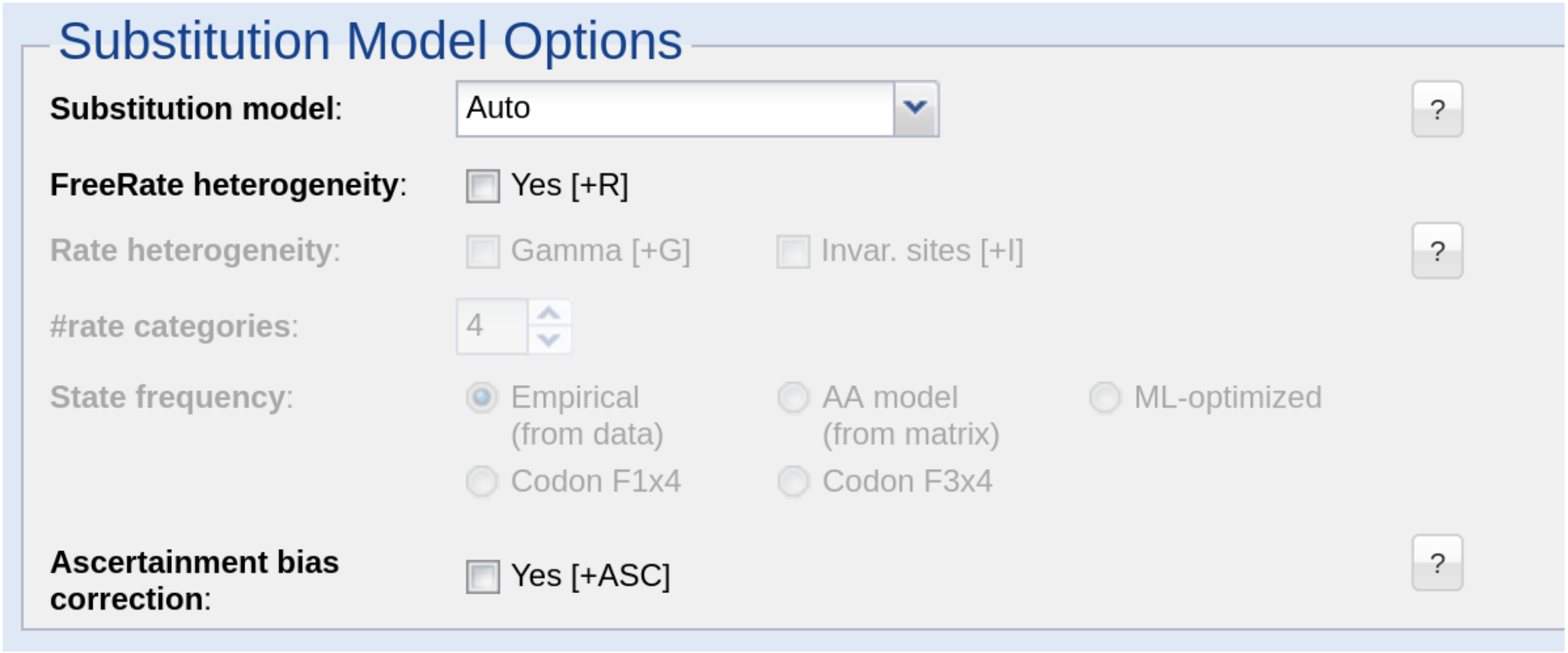
MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree



MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree



MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree

IQ-TREE Search Parameters

Perturbation strength: 0.5

IQ-TREE stopping rule: 100

Hide <

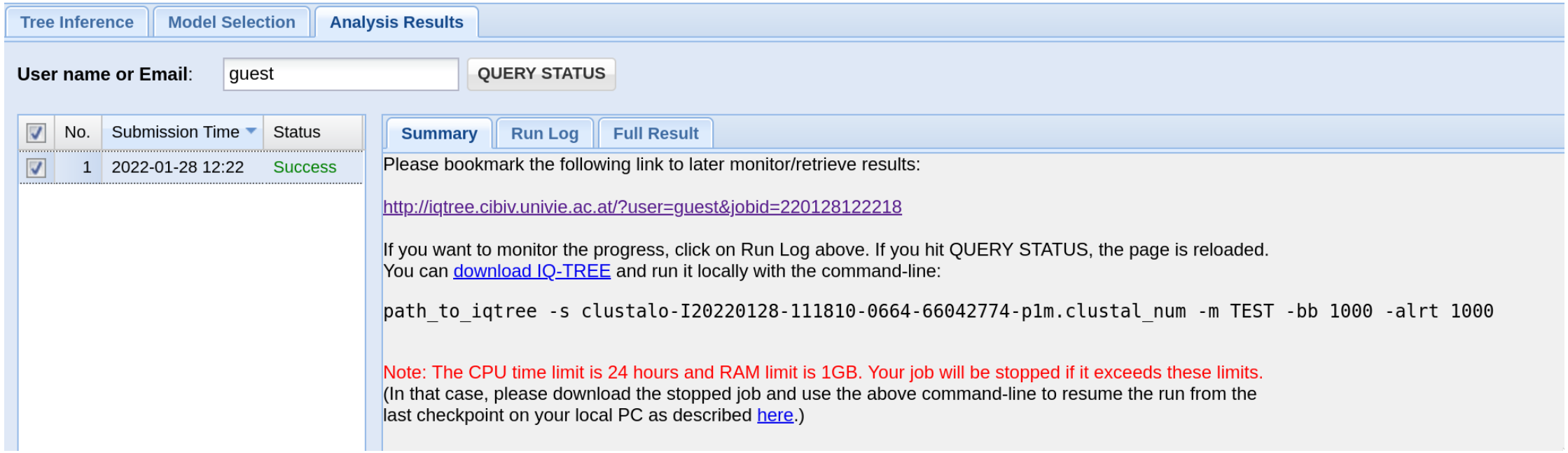
Help

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

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`.

MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree



MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree

```
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)
```

MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree

```
Tree in newick format:

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

MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree

```
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 (%)
```

MAX PERUTZ LABS VIENNA

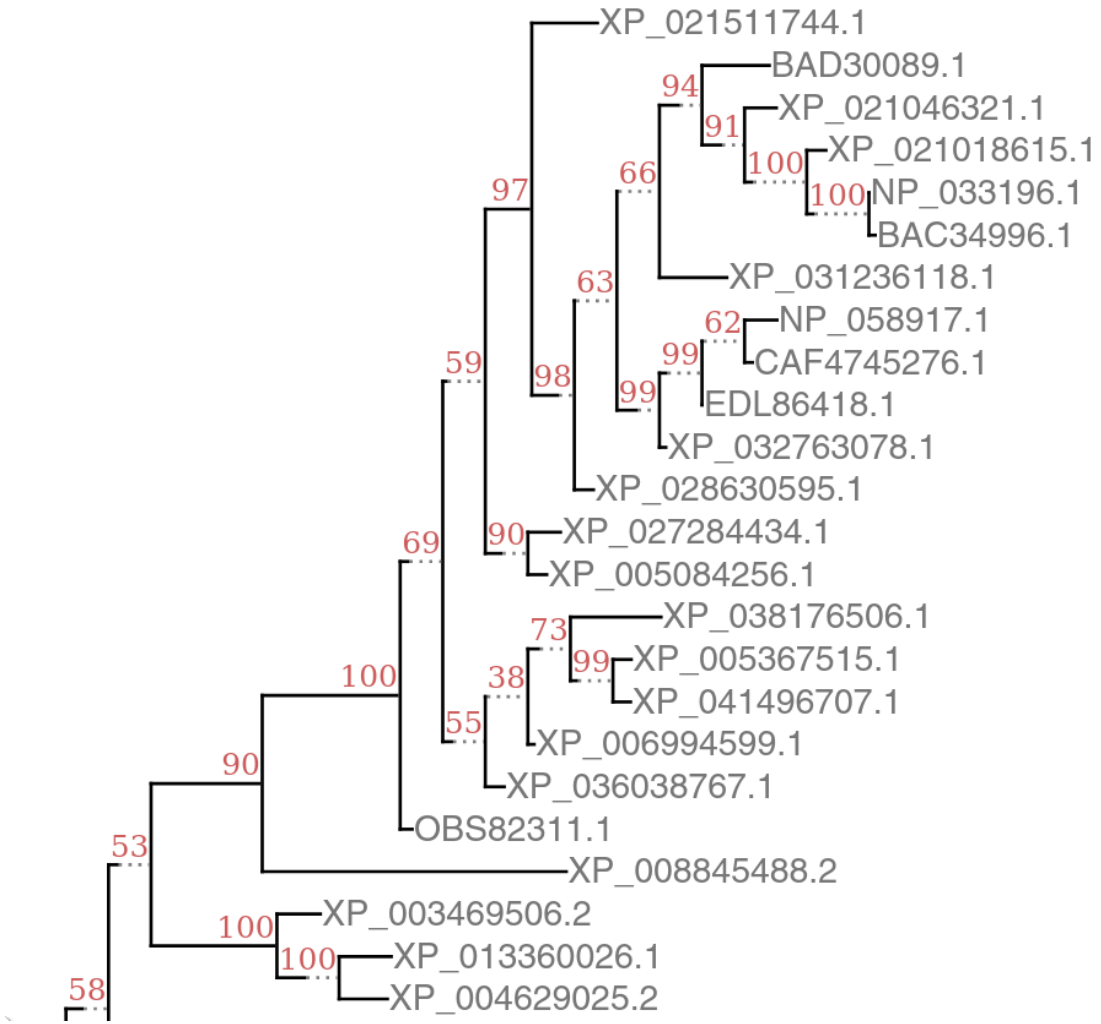
Building a Phylogenetic Tree

```
Consensus tree in newick format:

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

MAX PERUTZ LABS VIENNA

Building a Phylogenetic Tree



MAX PERUTZ LABS VIENNA

Ronja Reinhardt

Evolutionary sequence alignments

What do they tell us?

MAX PERUTZ LABS VIENNA

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?

MAX PERUTZ LABS VIENNA

How old is the gene I am working on?

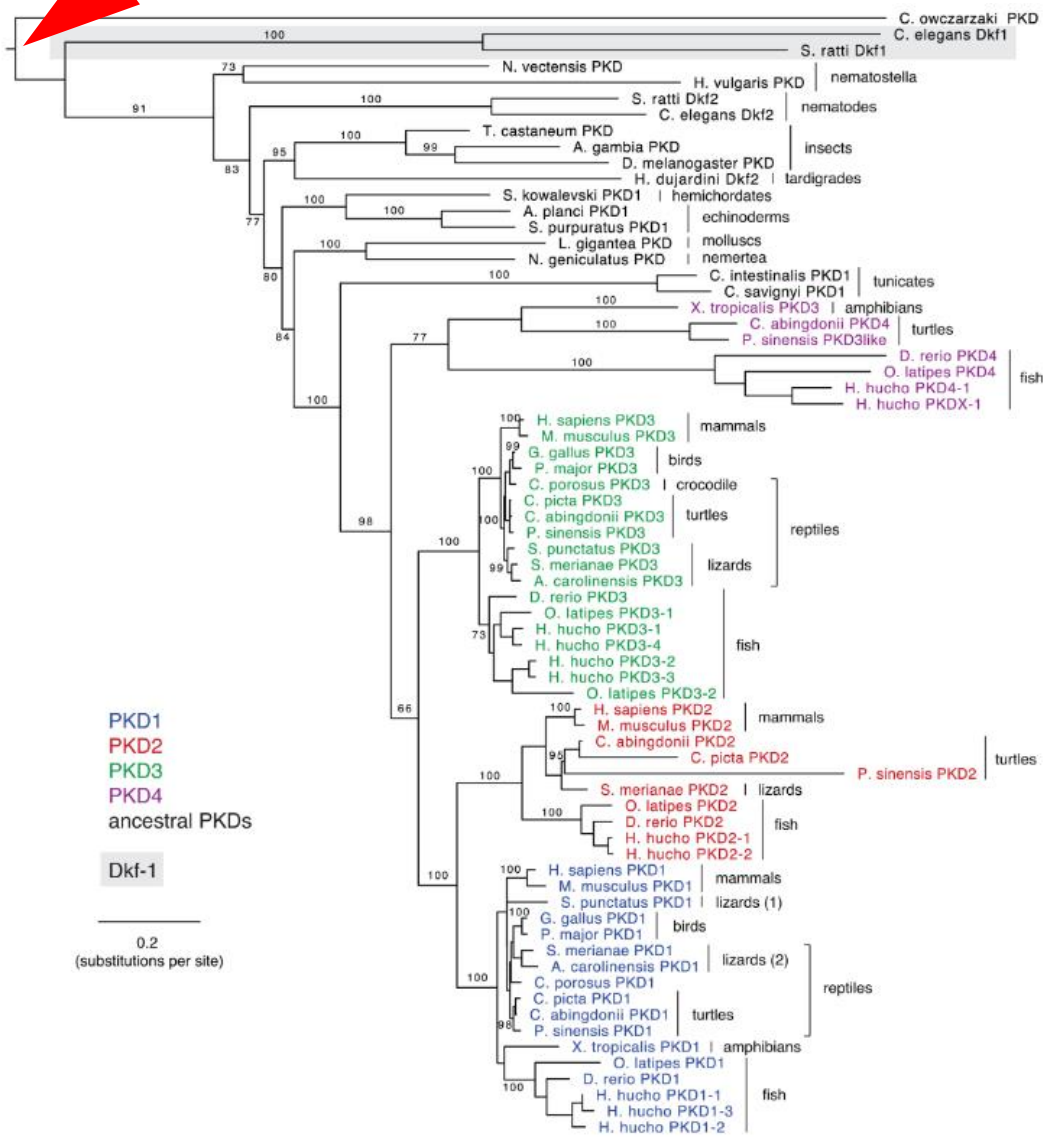
– Why am I curious about it?

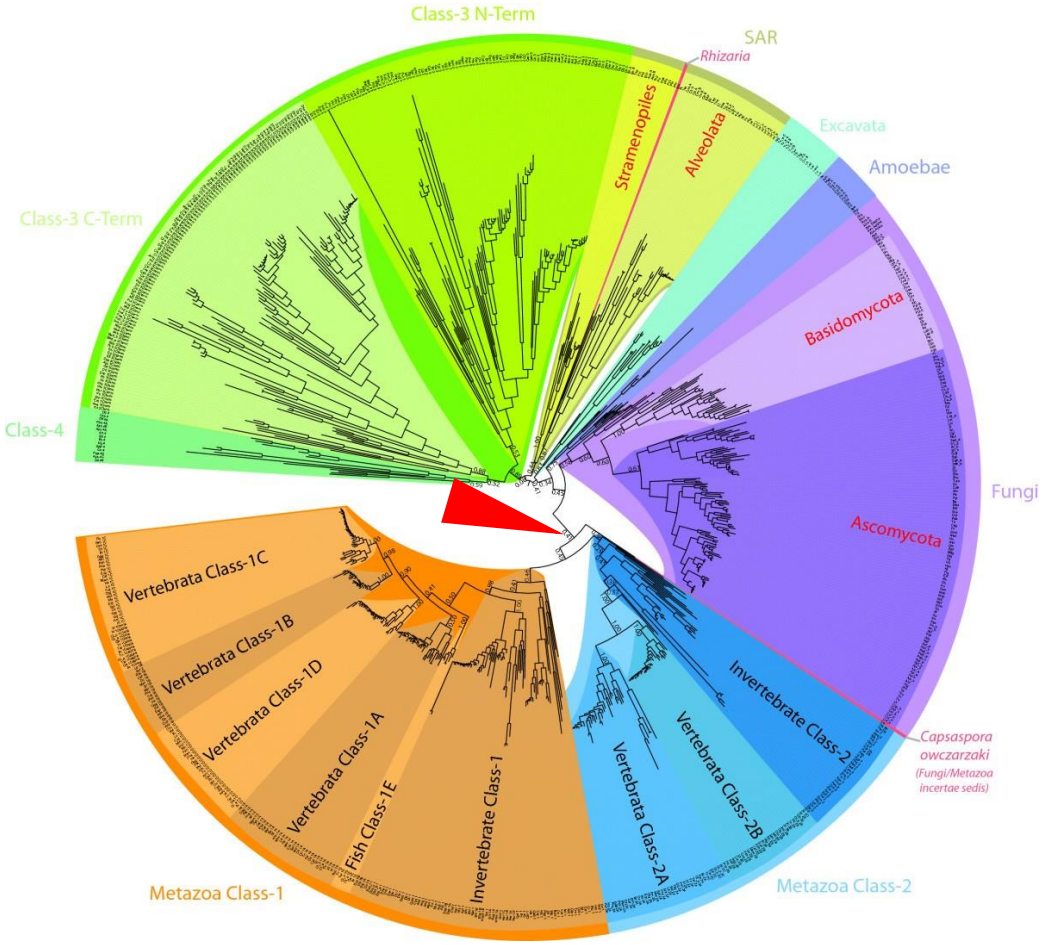
➢ What is the functional scope of my discovery?

- 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?

MAX PERUTZ LABS VIENNA

How old is the gene I am working on?

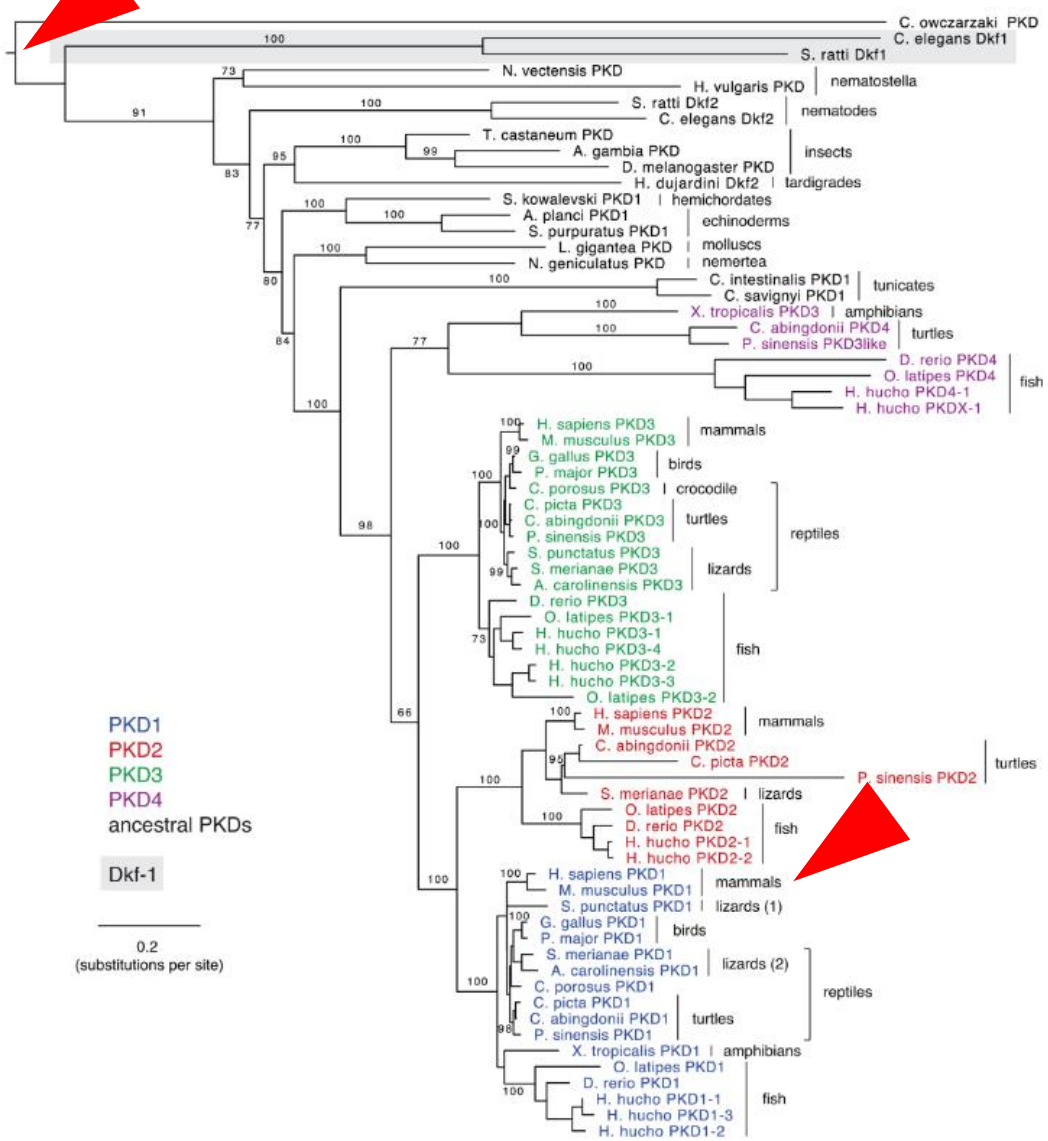


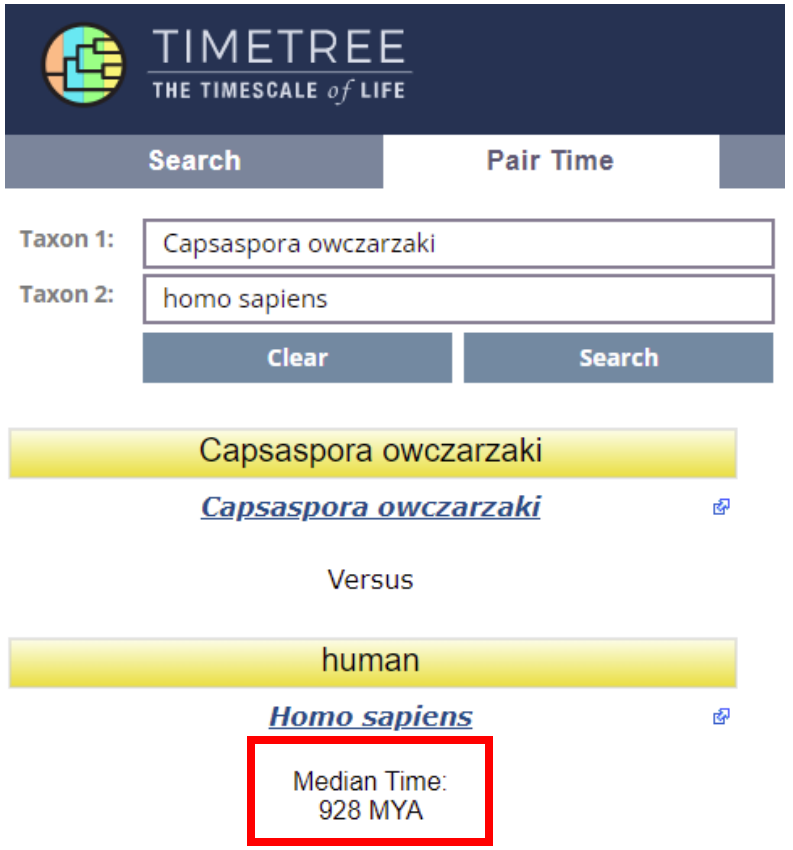


This protein/gene is conserved in **all animals**!

MAX PERUTZ LABS VIENNA

How old is the gene I am working on?





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

MAX PERUTZ LABS VIENNA

Which parts are of core importance?

– Why am I curious about this?

- Discovery of domains/motifs
- Construct design

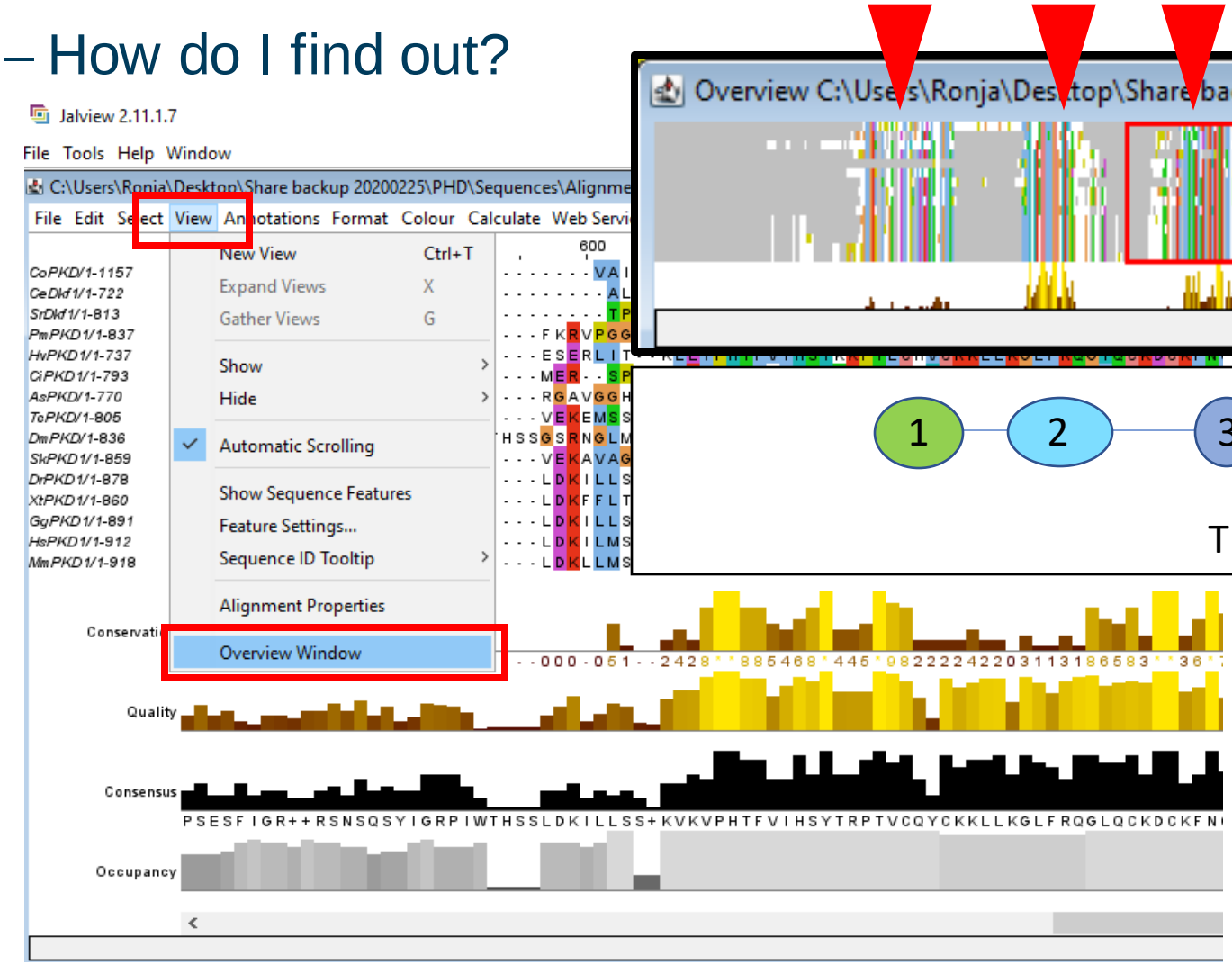
– How do I find out?

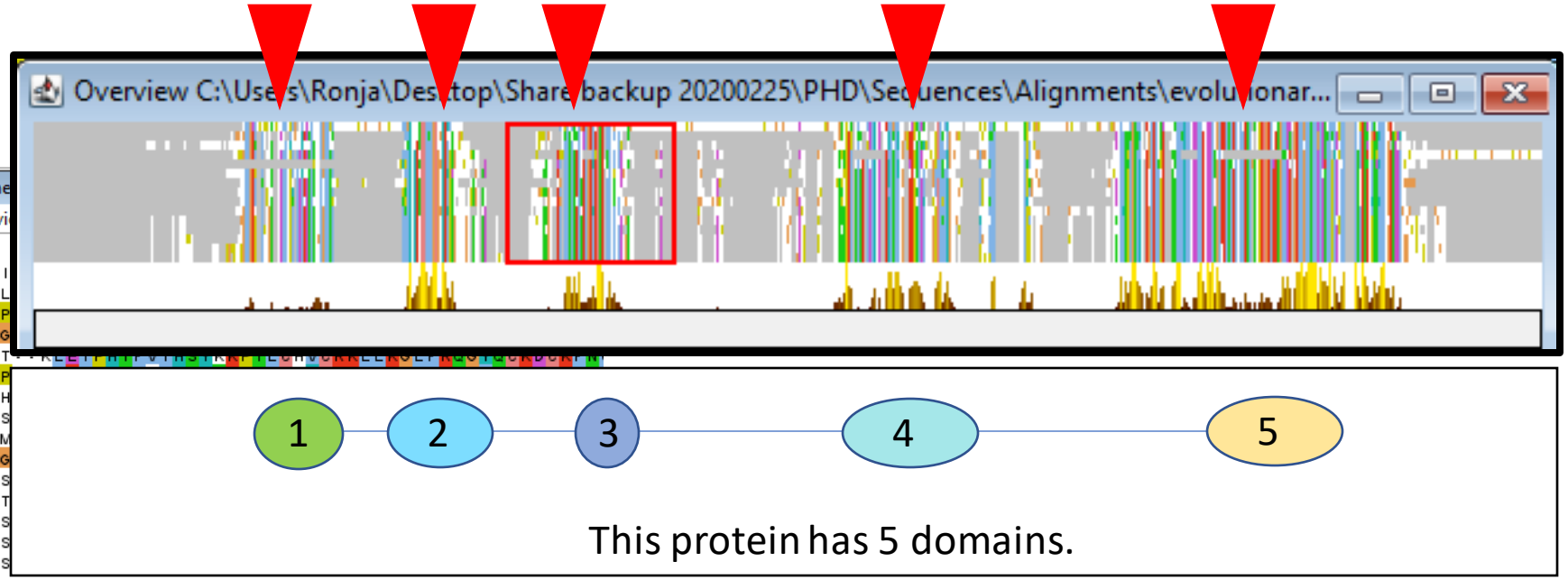
- Look at conserved parts in an alignment
- Check on Uniprot

MAX PERUTZ LABS VIENNA

Which parts are of core importance?

– How do I find out?

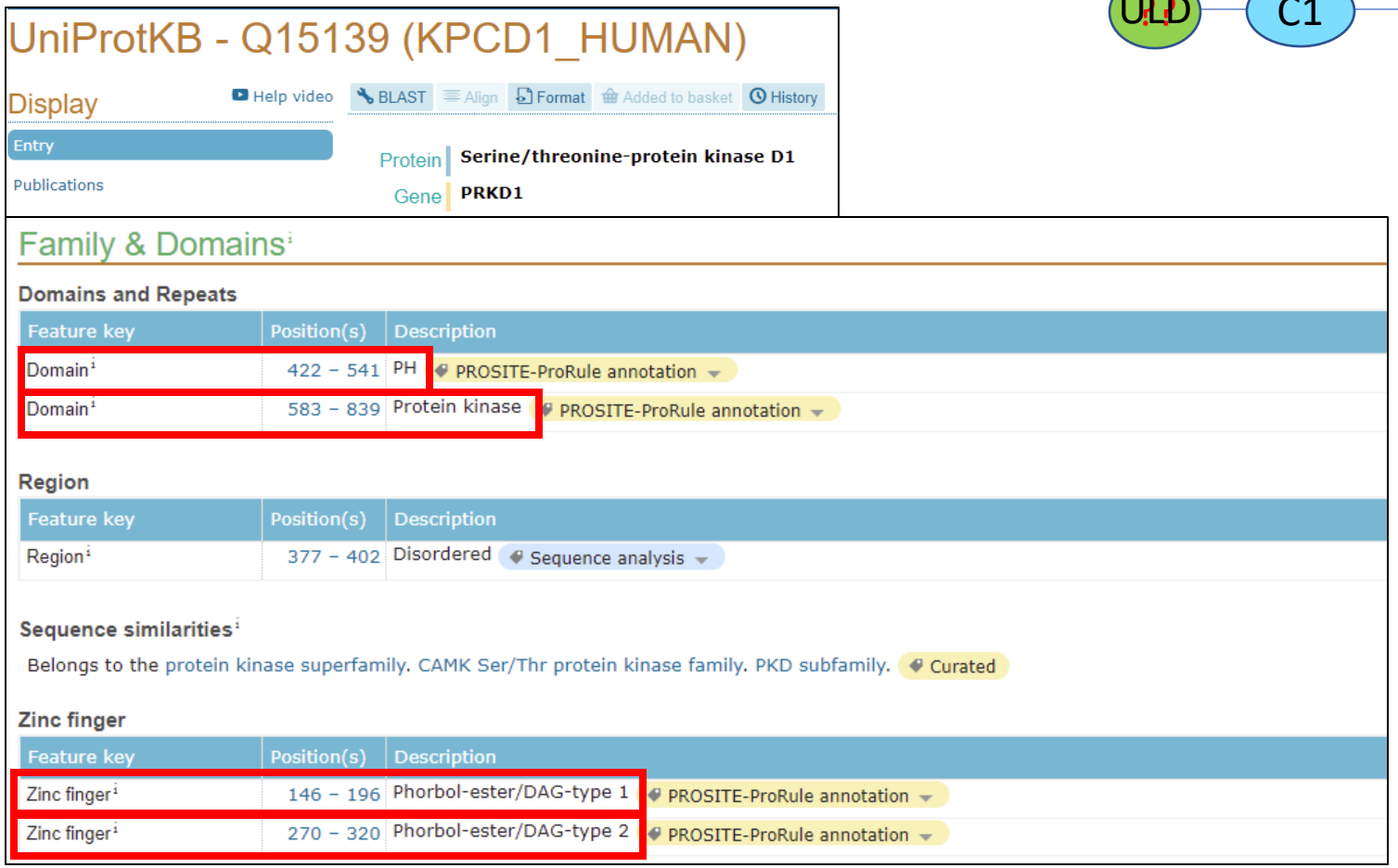


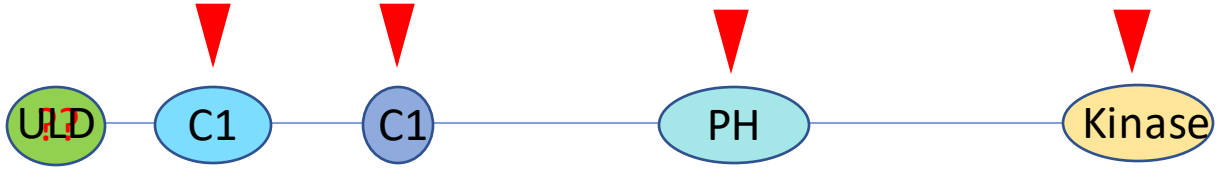


MAX PERUTZ LABS VIENNA

Which parts are of core importance?

– How do I find out?





This protein has 5 domains.

UniProt reports only 4.

New domain discovered ✓

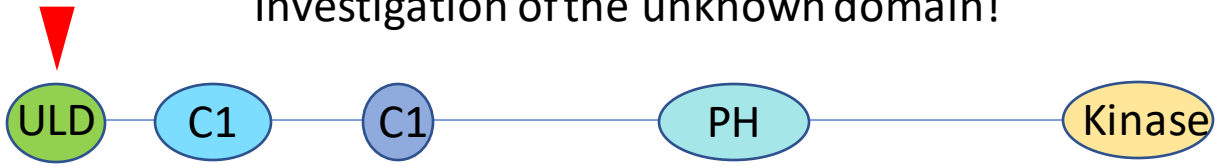
Elsner et al. 2019

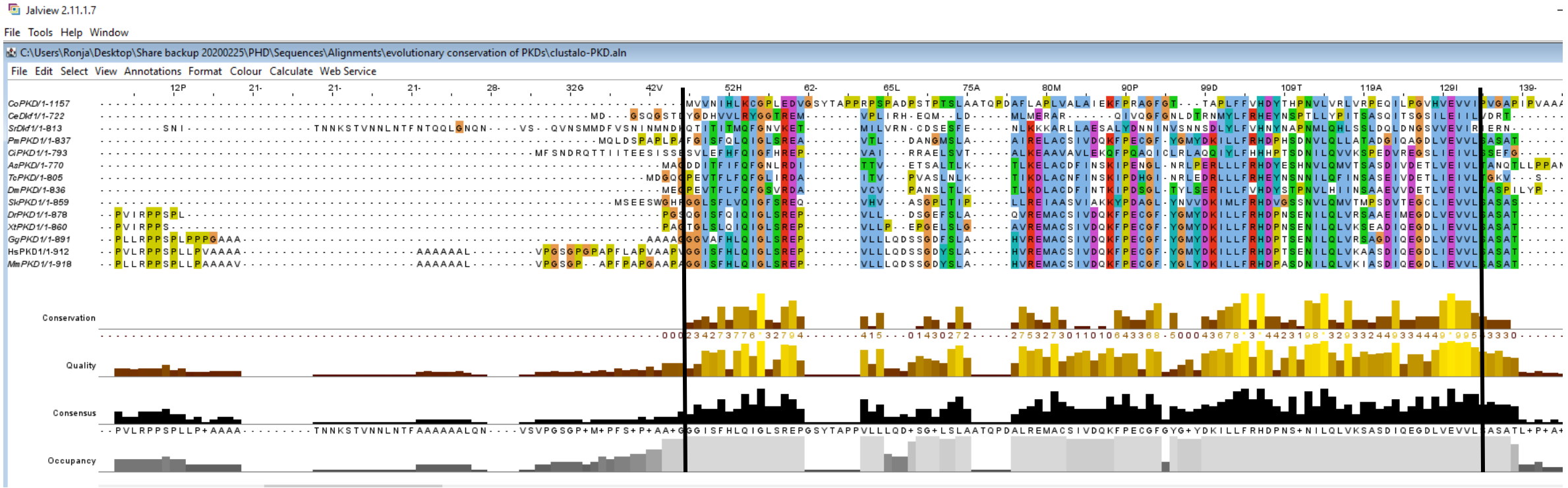
MAX PERUTZ LABS VIENNA

Which parts are of core importance?

- Construct design

Investigation of the unknown domain!





MAX PERUTZ LABS VIENNA

Which properties are conserved?

– Why am I curious about this?

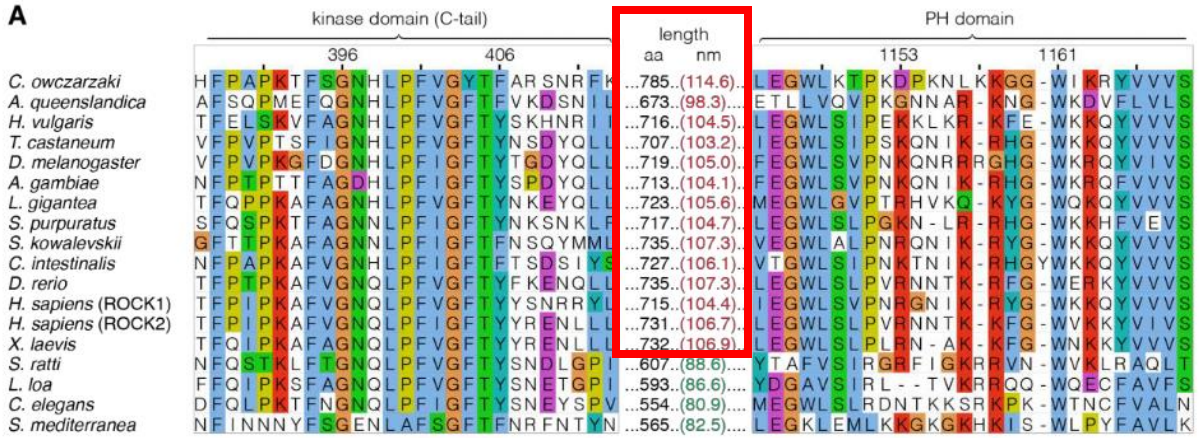
A conserved property is always an important property!

- Length
- Charge
- Hydrophobicity
- …

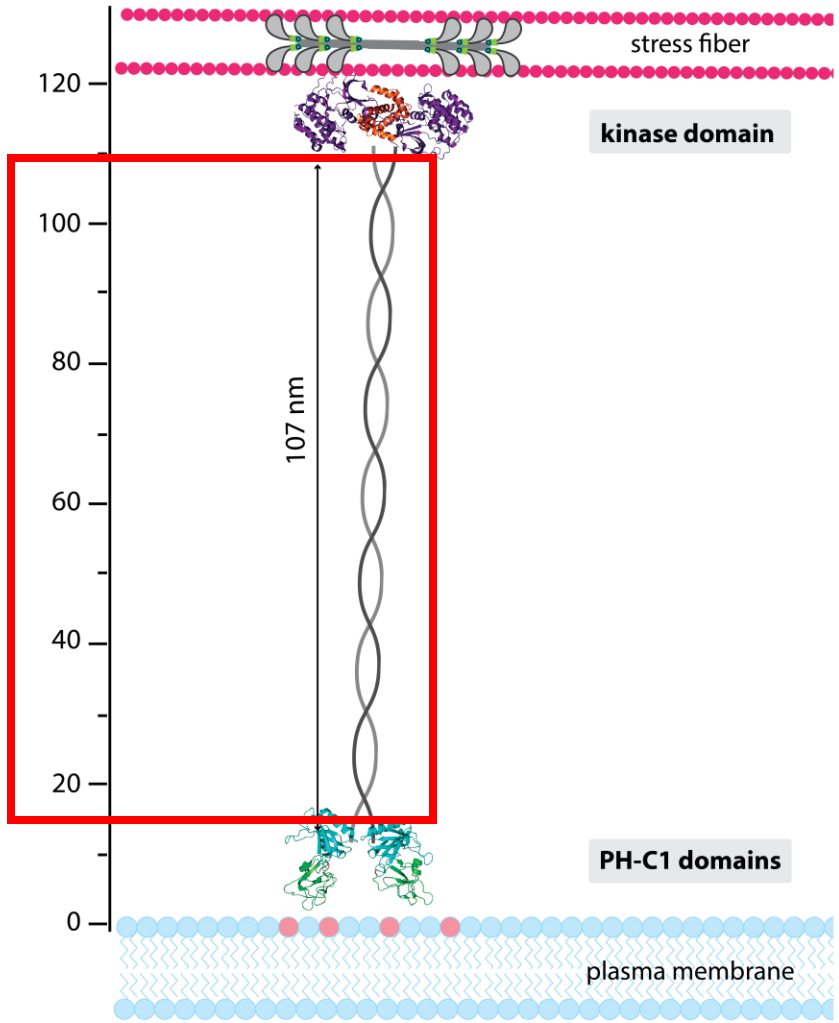
MAX PERUTZ LABS VIENNA

Which properties are important?

- Conserved length



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

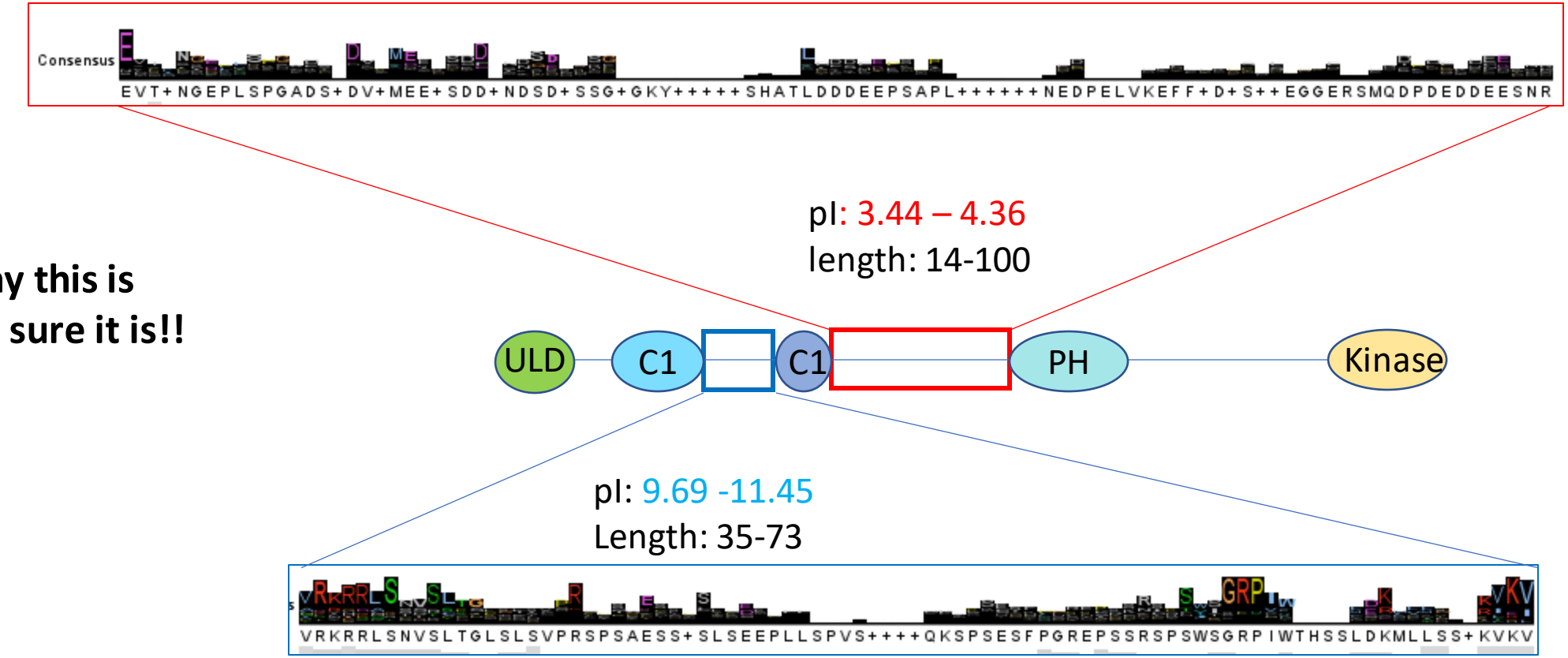


MAX PERUTZ LABS VIENNA

Which properties are important?

- Charge

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



MAX PERUTZ LABS VIENNA

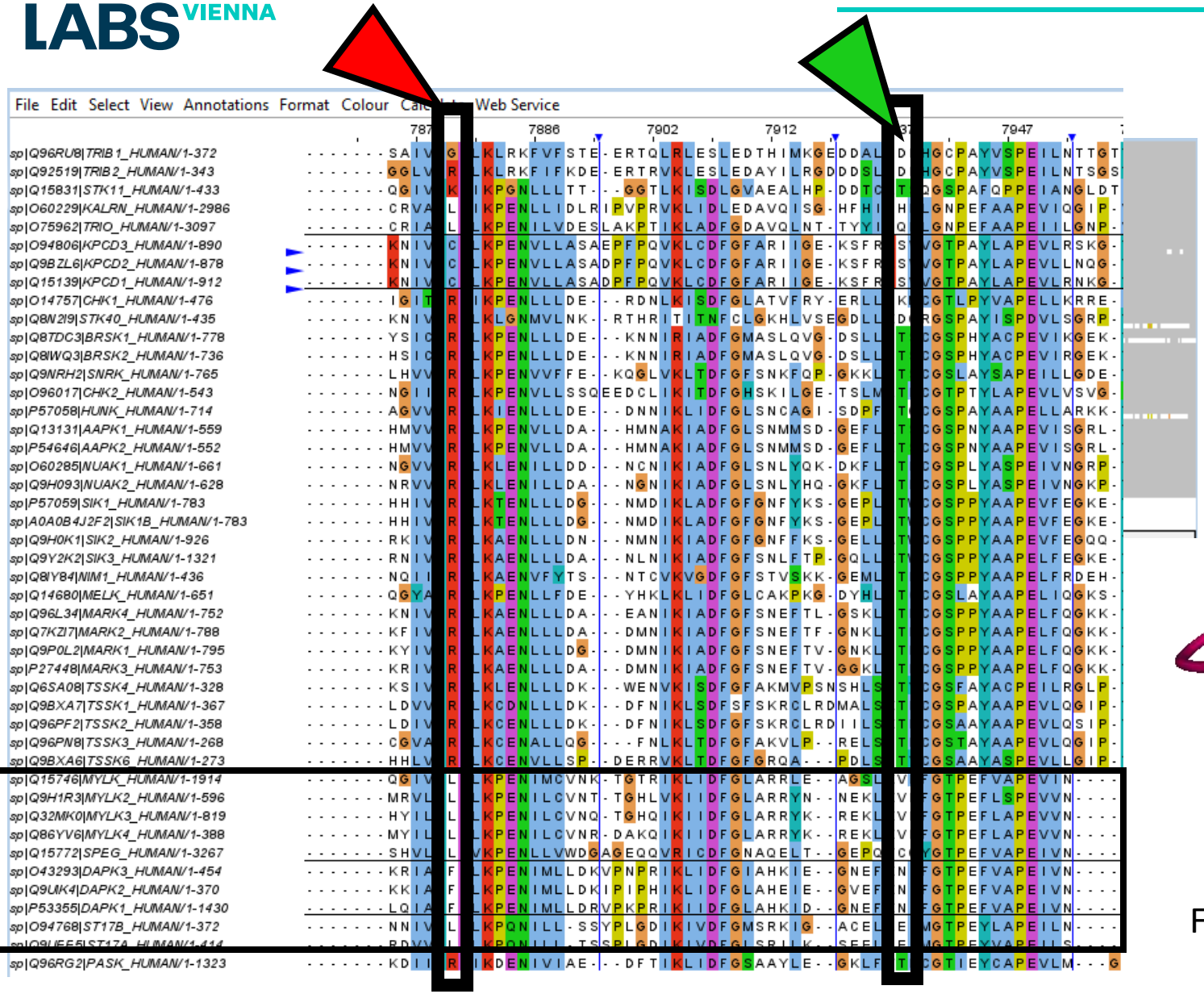
Which parts are dependent on each other?

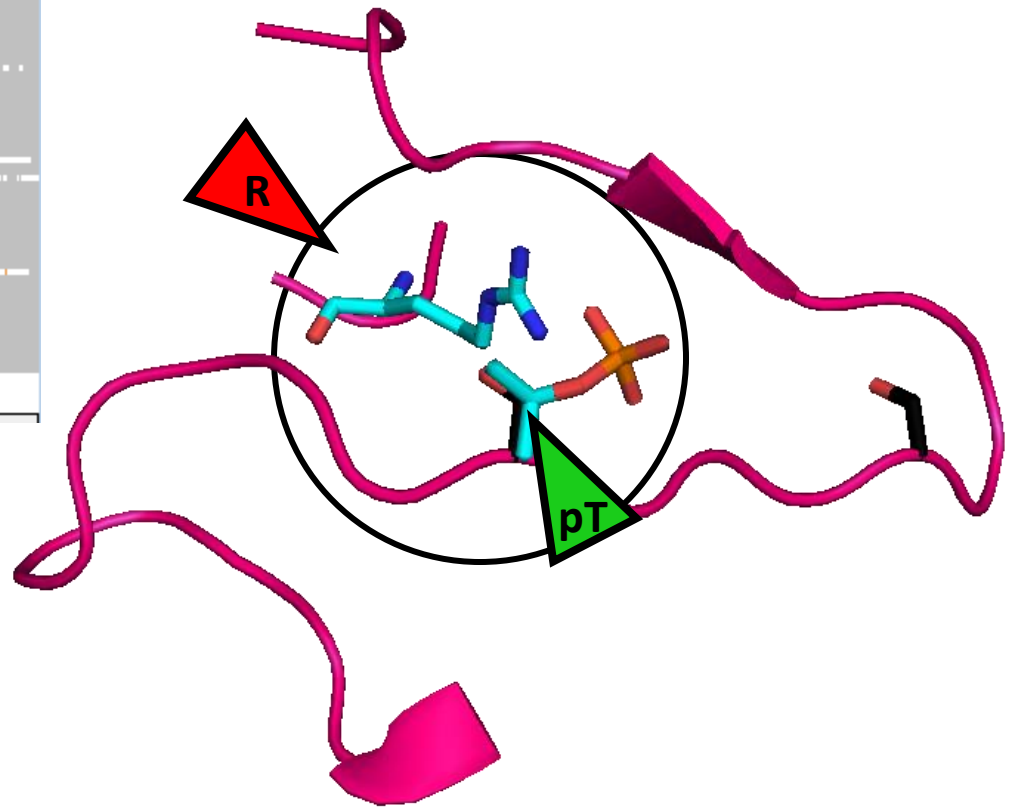
– Why am I curious about this?

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

MAX PERUTZ LABS VIENNA

Which parts are dependent on each other?

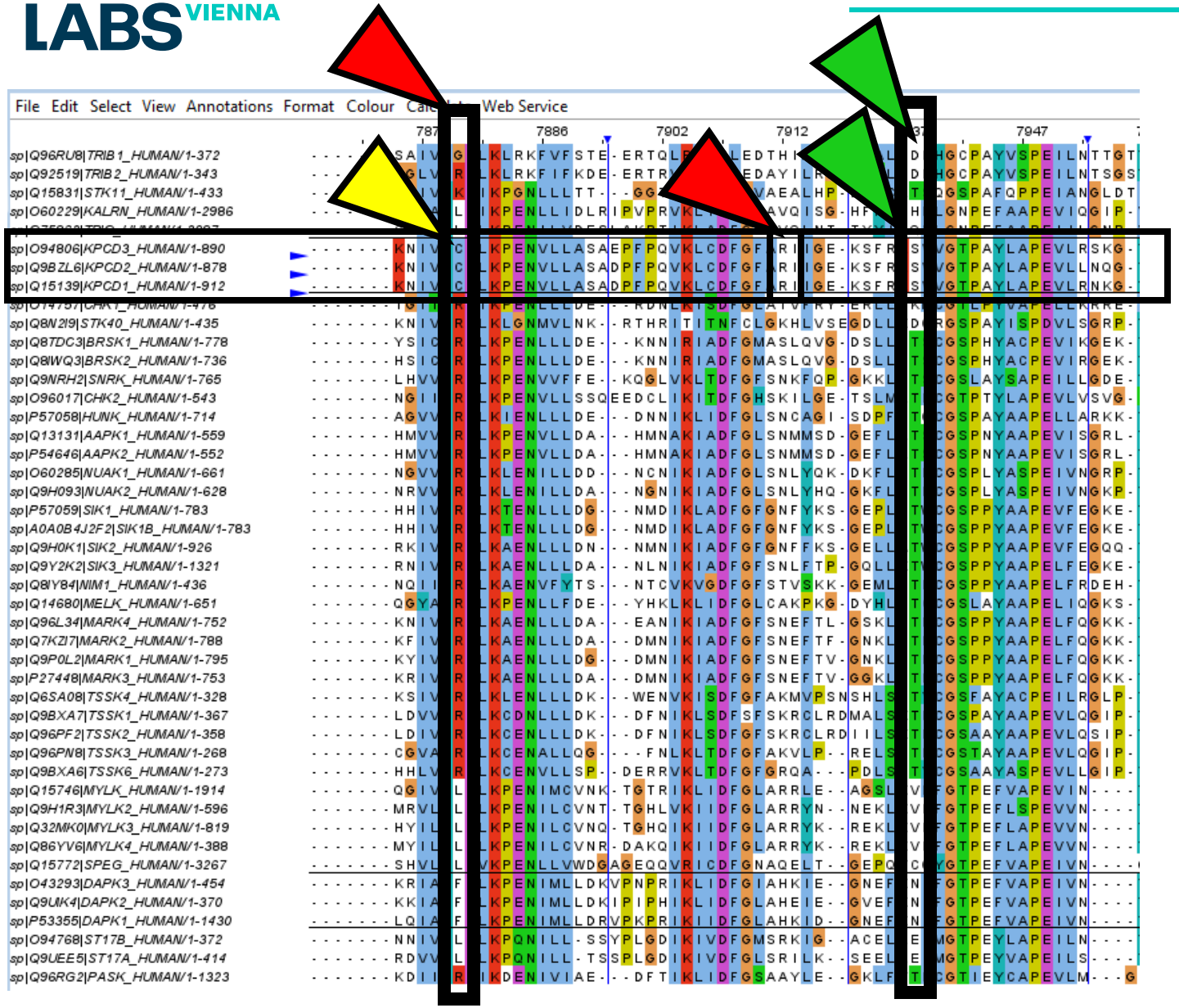


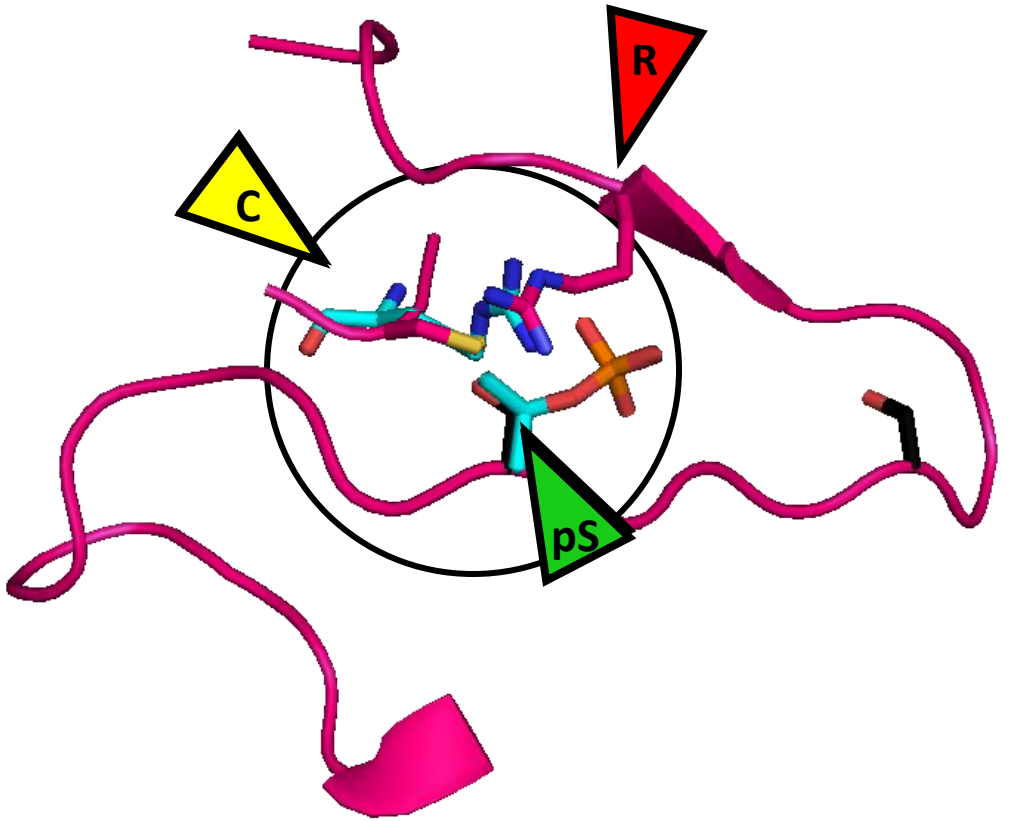


Functionally connected residues evolve together

MAX PERUTZ LABS VIENNA

Which parts are dependent on each other?





Important features don’t get lost in evolution!

MAX PERUTZ LABS VIENNA

Why are sequence alignments amazing?

- 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

MAX PERUTZ LABS VIENNA

Thank you for your attention!

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

Find us and ask ☺

MAX PERUTZ LABS VIENNA

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