UNIVERSITY OF
CAMBRIDGE

Bioinformatics for Protein Expression

Day 2: Tuesday 21st March

Today's talk

- How much can we learn and predict from sequence alone?

- How does this help us with our purifications?

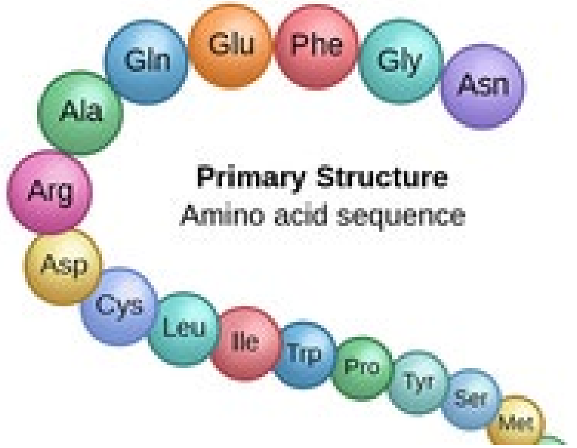
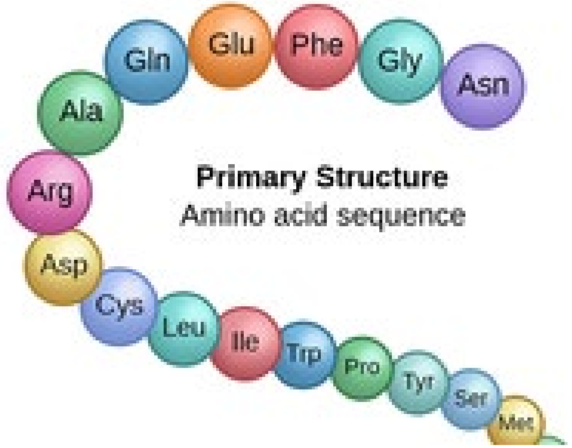


Image: Labster Theory

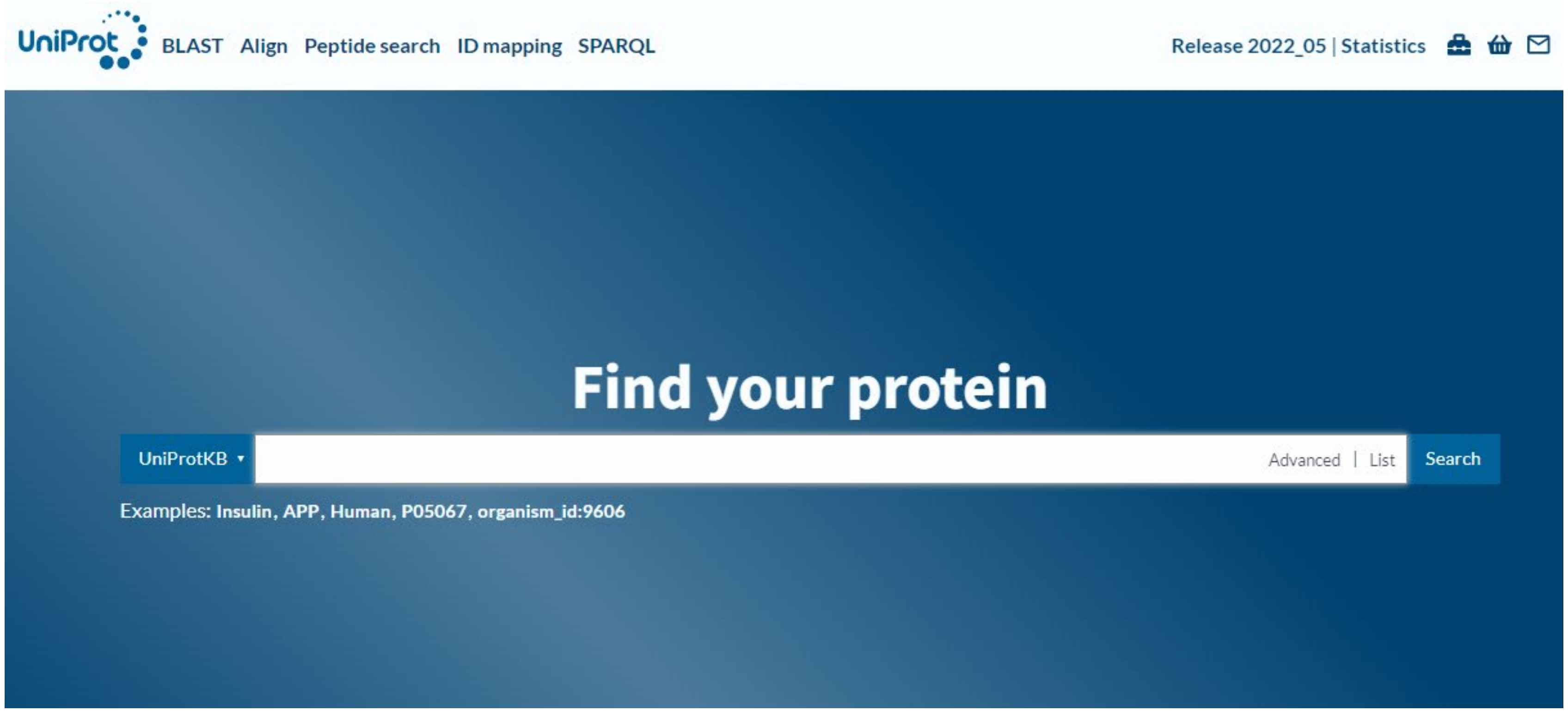
How can bioinformatics help me?

- In this lecture you will learn how (using just your protein sequence) you can:
  - Predict domains
  - Identify post translational modifications
  - Calculate the molecular mass
  - Determine the isoelectric point
  - And most importantly, learn how to get your “extinction co-efficients” so you can calculate protein concentration accurately
  - Identify distant homologs



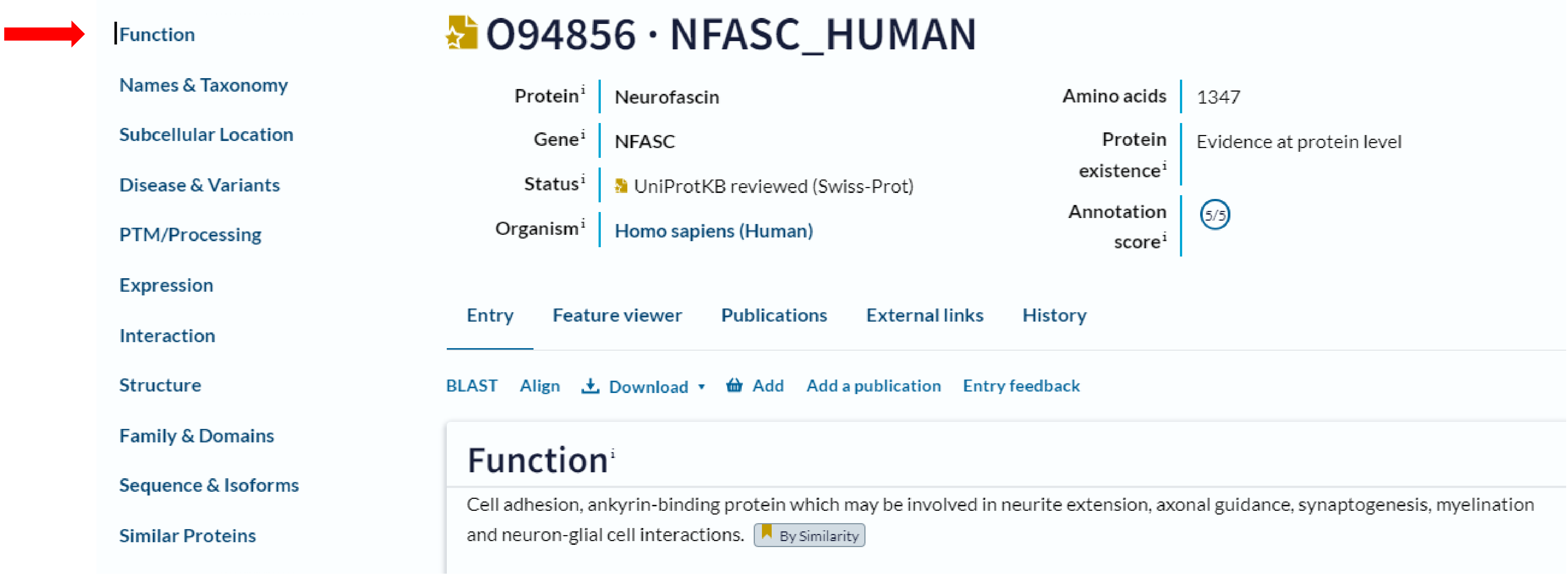
First useful resource - Uniprot

- Uniprot.org



Examples: Insulin, APP, Human, P05067, organism_id:9606

Uniprot – brief functional description



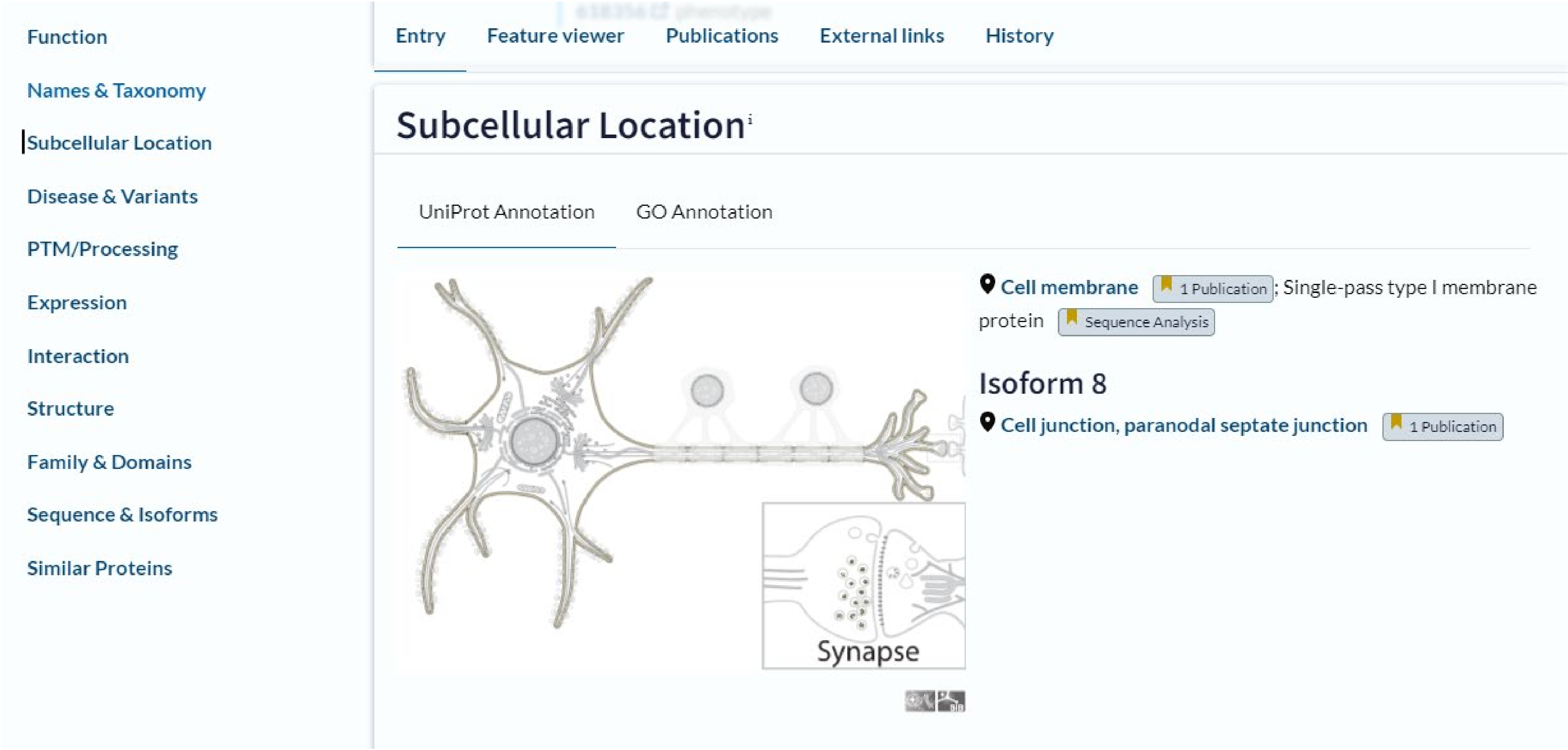
O94856 · NFASC_HUMAN

Protein	Neurofascin	Amino acids	1347
Gene	NFASC	Protein existence	Evidence at protein level
Status	UniProtKB reviewed (Swiss-Prot)	Annotation score	5/5
Organism	Homo sapiens (Human)		

Function

Cell adhesion, ankyrin-binding protein which may be involved in neurite extension, axonal guidance, synaptogenesis, myelination and neuron-glial cell interactions. By Similarity

Uniprot – subcellular localisation



- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoforms
- Similar Proteins

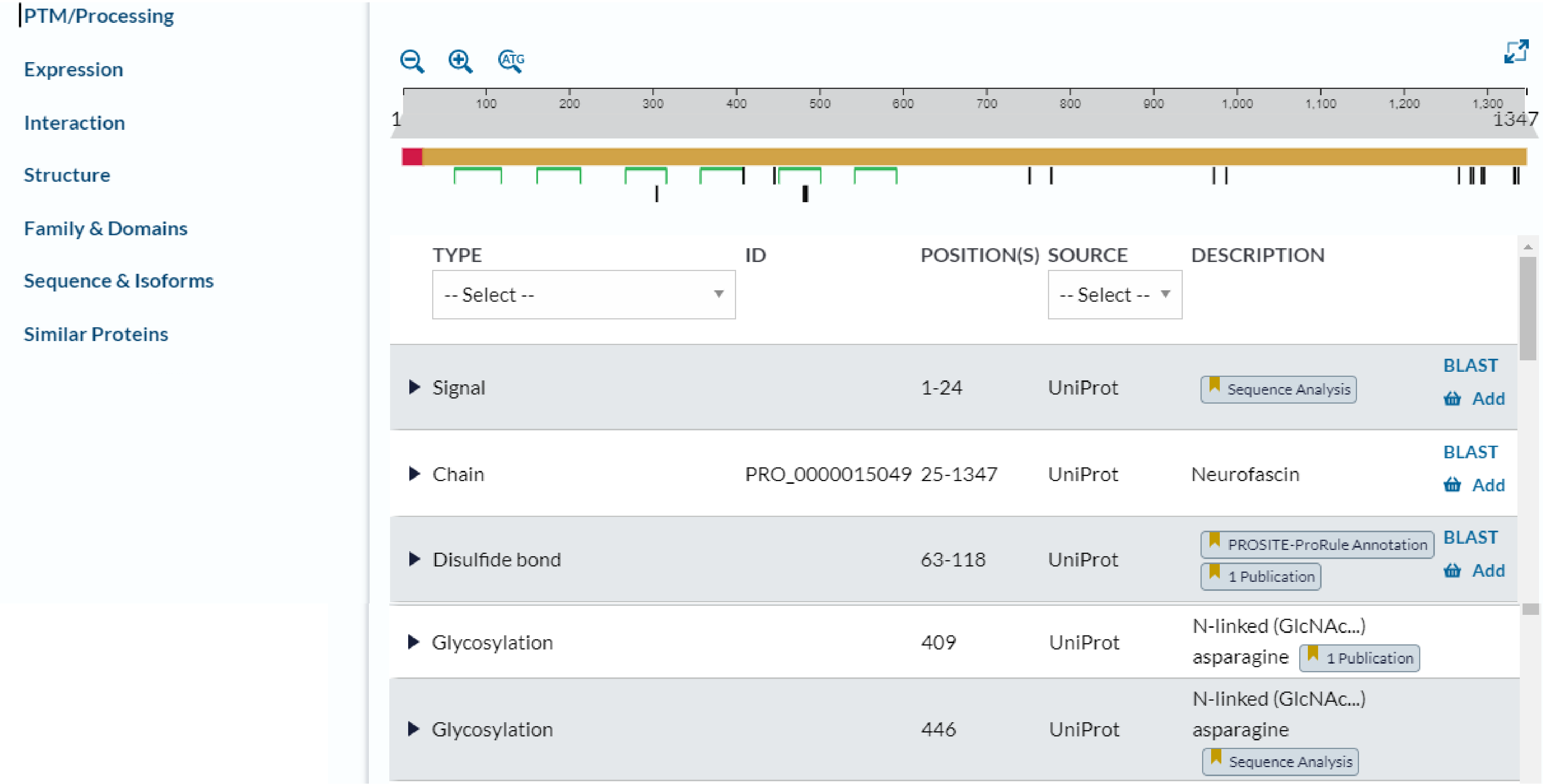
Subcellular Location

Cell membrane 1 Publication; Single-pass type I membrane protein Sequence Analysis

Isoform 8

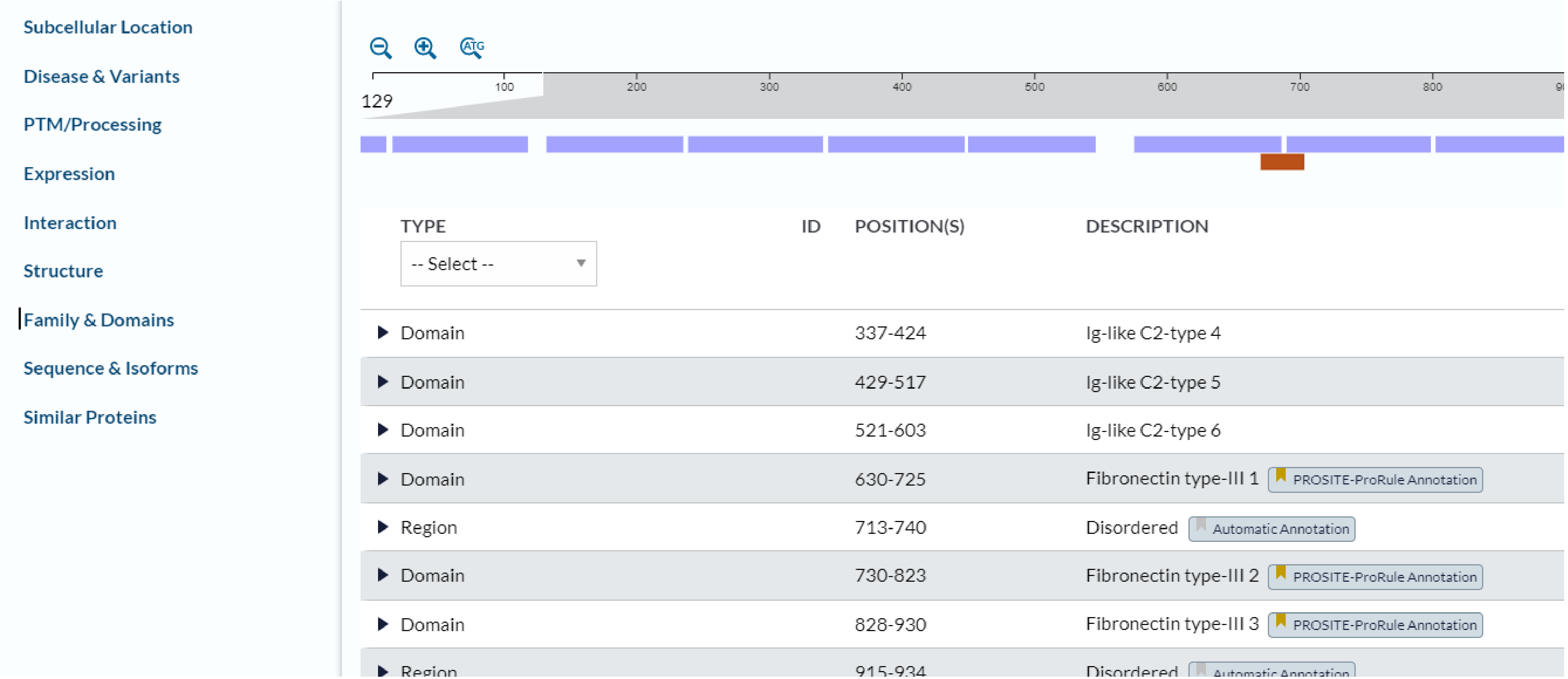
Cell junction, paranodal septate junction 1 Publication

Uniprot – post-translational modifications



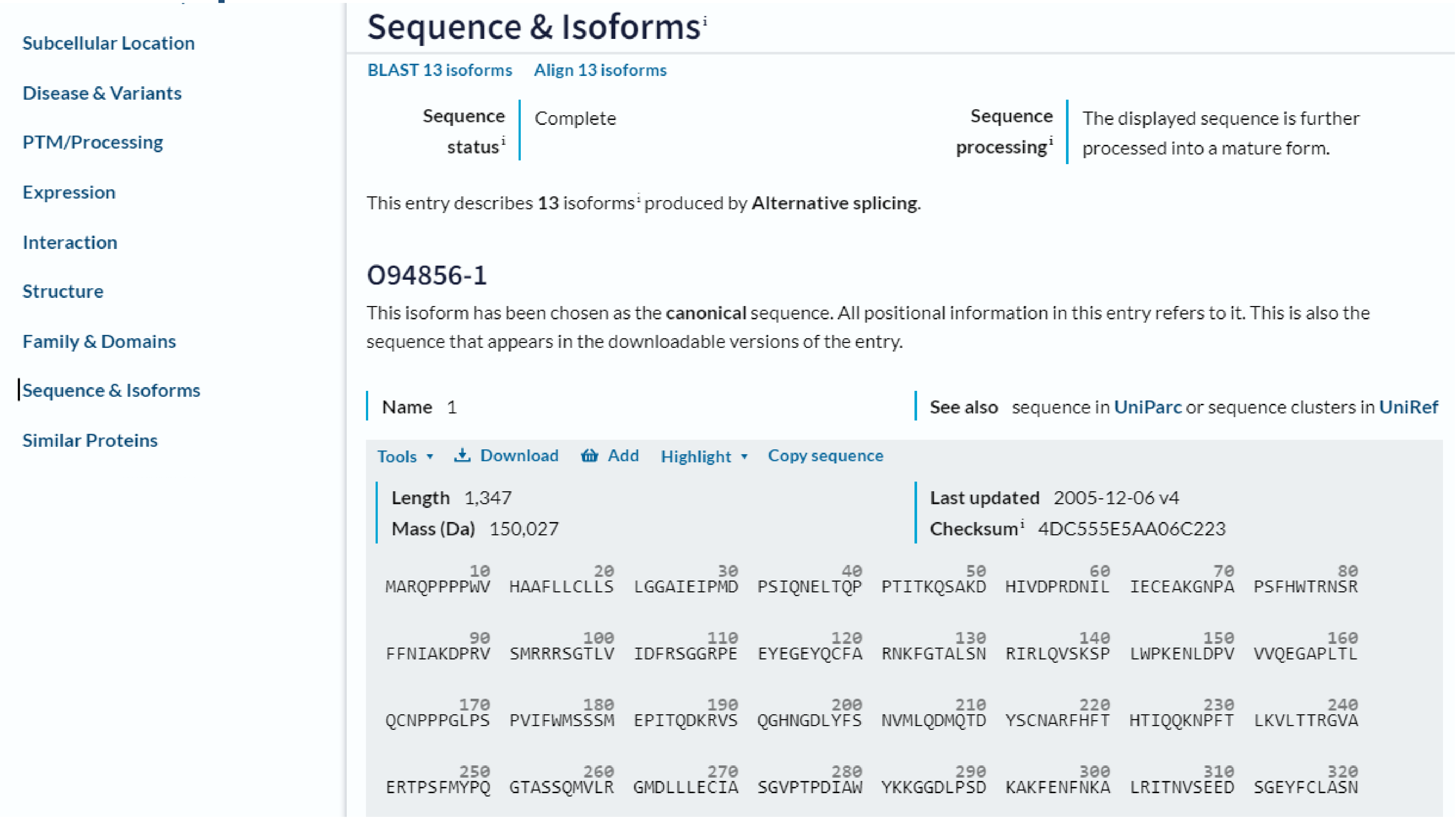
TYPE	ID	POSITION(S)	SOURCE	DESCRIPTION	
-- Select --			-- Select --		
▶ Signal		1-24	UniProt	Sequence Analysis	BLAST Add
▶ Chain	PRO_0000015049	25-1347	UniProt	Neurofascin	BLAST Add
▶ Disulfide bond		63-118	UniProt	PROSITE-ProRule Annotation 1 Publication	BLAST Add
▶ Glycosylation		409	UniProt	N-linked (GlcNAc...) asparagine 1 Publication	
▶ Glycosylation		446	UniProt	N-linked (GlcNAc...) asparagine Sequence Analysis	

Uniprot – Domain composition



TYPE -- Select --	ID	POSITION(S)	DESCRIPTION
▶ Domain		337-424	Ig-like C2-type 4
▶ Domain		429-517	Ig-like C2-type 5
▶ Domain		521-603	Ig-like C2-type 6
▶ Domain		630-725	Fibronectin type-III 1 PROSITE-ProRule Annotation
▶ Region		713-740	Disordered Automatic Annotation
▶ Domain		730-823	Fibronectin type-III 2 PROSITE-ProRule Annotation
▶ Domain		828-930	Fibronectin type-III 3 PROSITE-ProRule Annotation
▶ Region		915-934	Disordered Automatic Annotation

Uniprot – sequence data



Sequence & Isoforms[i]

BLAST 13 isoforms   Align 13 isoforms

Sequence status[i] Complete

Sequence processing[i] The displayed sequence is further processed into a mature form.

This entry describes **13** isoforms[i] produced by **Alternative splicing**.

O94856-1

This isoform has been chosen as the **canonical** sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Name 1

See also sequence in UniParc or sequence clusters in UniRef

Length 1,347

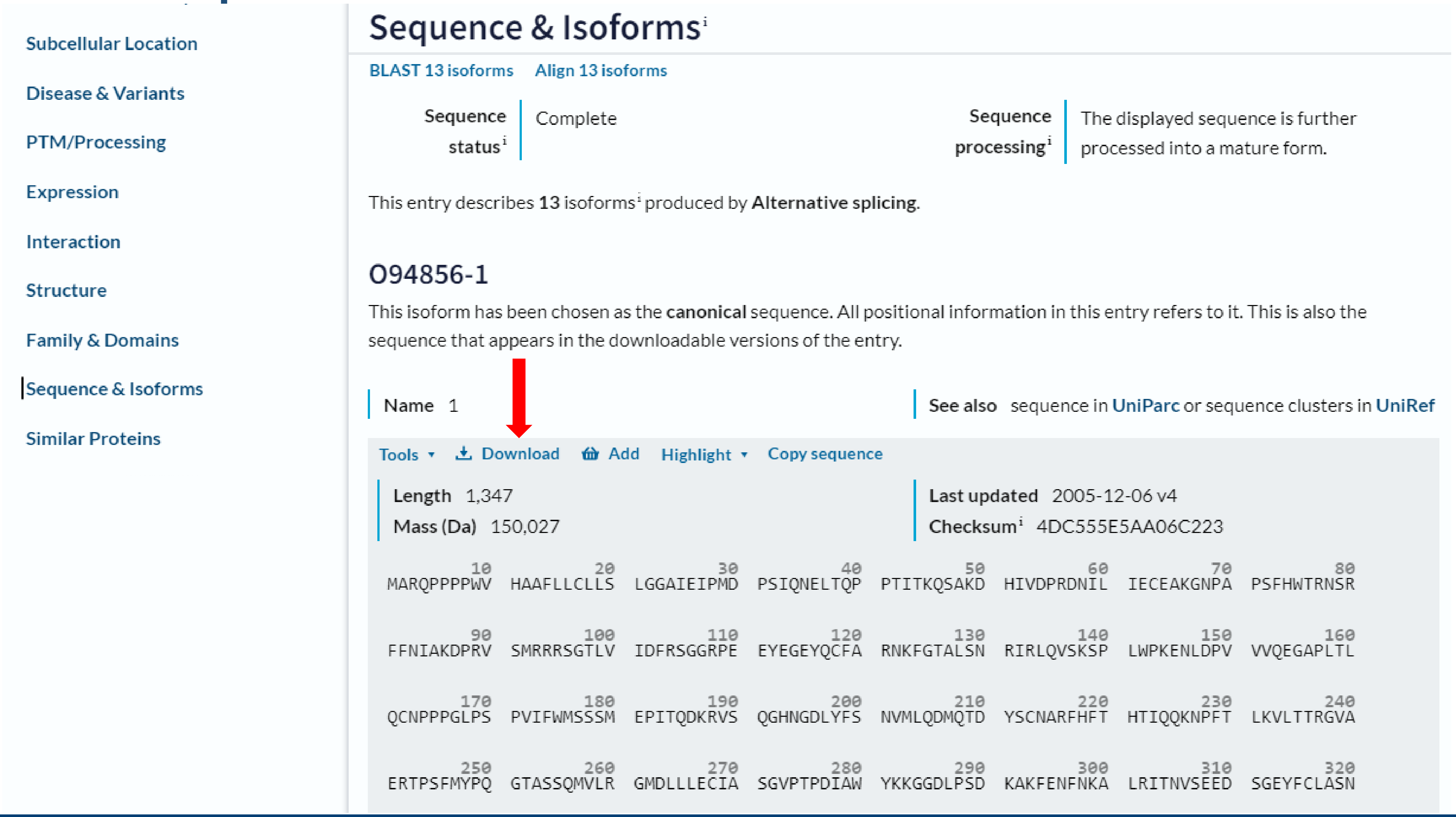
Mass (Da) 150,027

Last updated 2005-12-06 v4

Checksum[i] 4DC555E5AA06C223

```
        10         20         30         40         50         60         70         80
MARQPPPPWV HAAFLLCLLS LGGAIEIPMD PSIQNELTQP PTITKQSAKD HIVDPRDNIL IECEAKGNPA PSFHWTRNSR

        90        100        110        120        130        140        150        160
FFNIAKDPRV SMRRRSGTLV IDFRSGGRPE EYEGEYQCFA RNKFGTALSN RIRLQVSKSP LWPKENLDPV VVQEGAPLTL

       170        180        190        200        210        220        230        240
QCNPPPGLPS PVIFWMSSSM EPITQDKRVS QGHNGDLYFS NVMLQDMQTD YSCNARFHFT HTIQQKNPFT LKVLTTRGVA

       250        260        270        280        290        300        310        320
ERTPSFMYPQ GTASSQMVLR GMDLLLECIA SGVPTPDIAW YKKGGDLPSD KAKFENFNKA LRITNVSEED SGEYFCLASN
```

Uniprot – sequence data



Sequence & Isoforms

Sequence status	Complete
Sequence processing	The displayed sequence is further processed into a mature form.

This entry describes **13** isoforms produced by **Alternative splicing**.

O94856-1

This isoform has been chosen as the **canonical** sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Name 1

See also sequence in UniParc or sequence clusters in UniRef

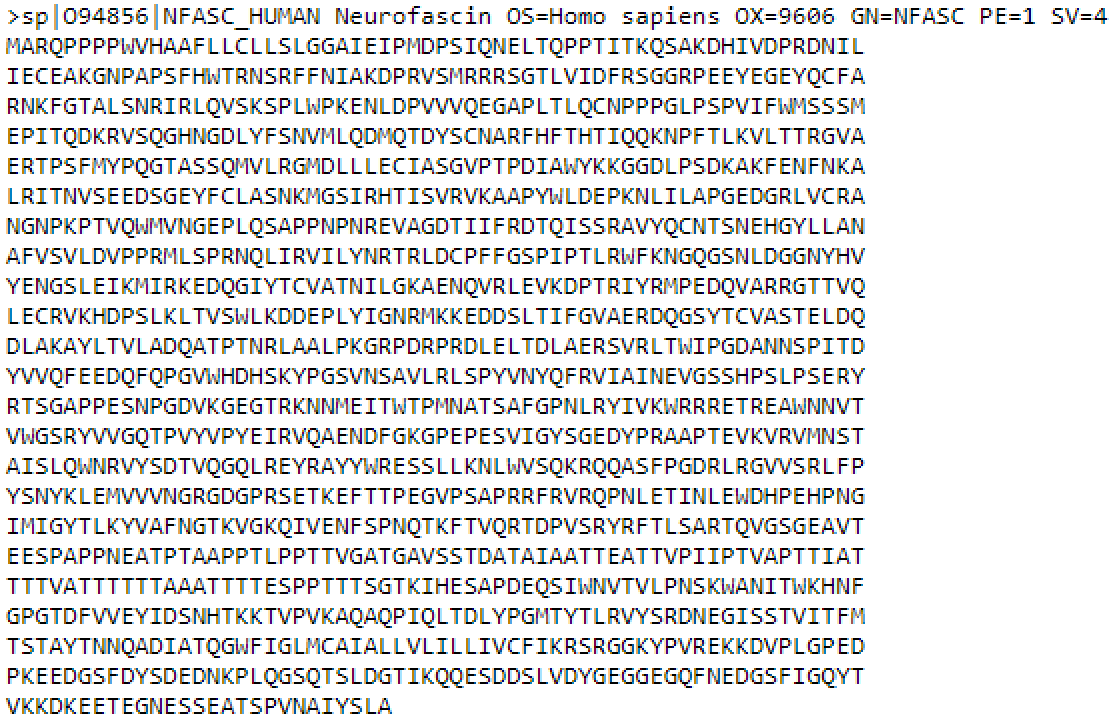
Length 1,347
Mass (Da) 150,027

Last updated 2005-12-06 v4
Checksum 4DC555E5AA06C223

```
        10         20         30         40         50         60         70         80
MARQPPPPWV HAAFLLCLLS LGGAIEIPMD PSIQNELTQP PTITKQSAKD HIVDPRDNIL IECEAKGNPA PSFHWTRNSR

        90        100        110        120        130        140        150        160
FFNIAKDPRV SMRRRSGTLV IDFRSGGRPE EYEGEYQCFA RNKFGTALSN RIRLQVSKSP LWPKENLDPV VVQEGAPLTL

       170        180        190        200        210        220        230        240
QCNPPPGLPS PVIFWMSSSM EPITQDKRVS QGHNGDLYFS NVMLQDMQTD YSCNARFHFT HTIQQKNPFT LKVLTTRGVA

       250        260        270        280        290        300        310        320
ERTPSFMYPQ GTASSQMVLR GMDLLLECIA SGVPTPDIAW YKKGGDLPSD KAKFENFNKA LRITNVSEED SGEYFCLASN
```

Uniprot – sequence formats

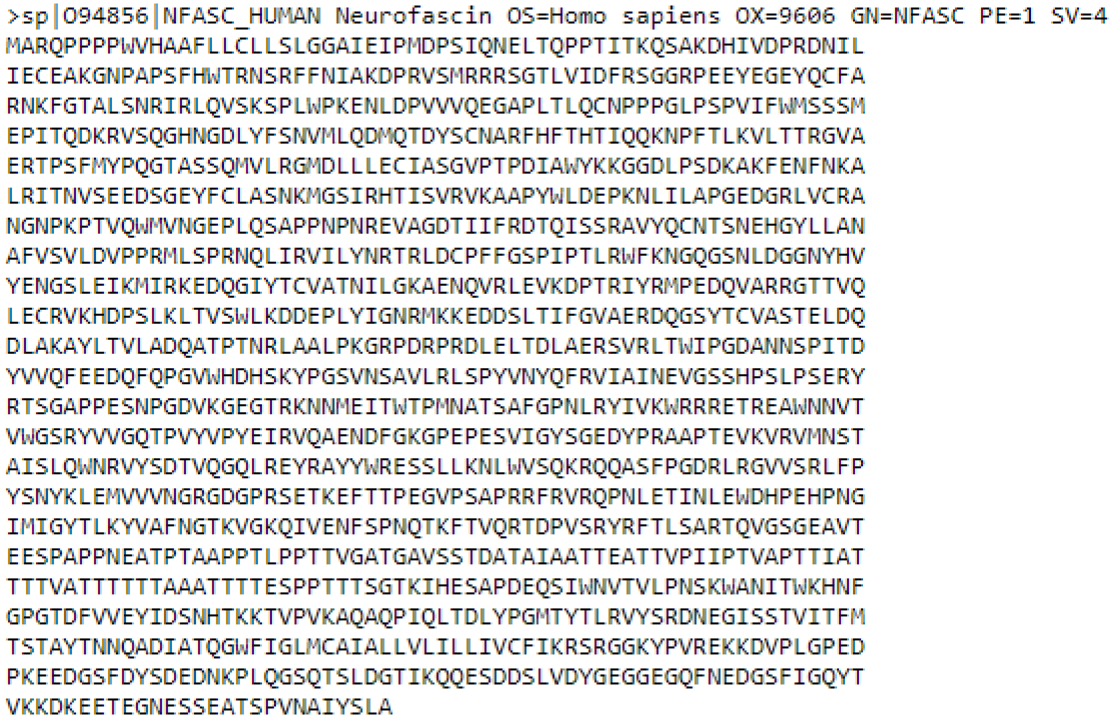
- FASTA format



```
>sp|O94856|NFASC_HUMAN Neurofascin OS=Homo sapiens OX=9606 GN=NFASC PE=1 SV=4
MARQPPPPWVHAAFLLCLLSLGGAIEIPMDPSIQNELTQPPTITKQSAKDHIVDPRDNIL
IECEAKGNPAPSFHWTRNSRFFNIAKDPRVSMRRRSGTLVIDFRSGGRPEEYEGEYQCFA
RNKFGTALSNRIRLQVSKSPLWPKENLDPVVVQEGAPLTLQCNPPPGLPSPVIFWMSSSM
EPITQDKRVSQGHNGDLYFSNVMLQDMQTDYSCNARFHFTHTIQQKNPFTLKVLTTRGVA
ERTPSFMYPQGTASSQMVLRGMDLLLECIASGVPTPDIAWYKKGGDLPSDKAKFENFNKA
LRITNVSEEDSGEYFCLASNKMGSIRHTISVRVKAAPYWLDEPKNLILAPGEDGRLVCRA
NGNPKPTVQWMVNGEPLQSAPPNPNREVAGDTIIFRDTQISSRAVYQCNTSNEHGYLLAN
AFVSVLDVPPRMLSPRNQLIRVILYNRTRLDCPFFGSPIPTLRWFKNGQGSNLDGGNYHV
YENGSLEIKMIRKEDQGIYTCVATNILGKAENQVRLEVKDPTRIYRMPEDQVARRGTTVQ
LECRVKHDPSLKLTVSWLKDDEPLYIGNRMKKEDDSLTIFGVAERDQGSYTCVASTELDQ
DLAKAYLTVLADQATPTNRLAALPKGRPDRPRDLELTDLAERSVRLTWIPGDANNSPITD
YVVQFEEDQFQPGVWHDHSKYPGSVNSAVLRLSPYVNYQFRVIAINEVGSSHPSLPSERY
RTSGAPPESNPGDVKGEGTRKNNMEITWTPMNATSAFGPNLRYIVKWRRRETREAWNNVT
VWGSRYVVGQTPVYVPYEIRVQAENDFGKGPEPESVIGYSGEDYPRAAPTEVKVRVMNST
AISLQWNRVYSDTVQGQLREYRAYYWRESSLLKNLWVSQKRQQASFPGDRLRGVVSRLFP
YSNYKLEMVVVNGRGDGPRSETKEFTTPEGVPSAPRRFRVRQPNLETINLEWDHPEHPNG
IMIGYTLKYVAFNGTKVGKQIVENFSPNQTKFTVQRTDPVSRYRFTLSARTQVGSGEAVT
EESPAPPNEATPTAAPPTLPPTTVGATGAVSSTDATAIAATTEATTVPIIPTVAPTTIAT
TTTVATTTTTTAAATTTTESPPTTTSGTKIHESAPDEQSIWNVTVLPNSKWANITWKHNF
GPGTDFVVEYIDSNHTKKTVPVKAQAQPIQLTDLYPGMTYTLRVYSRDNEGISSTVITFM
TSTAYTNNQADIATQGWFIGLMCAIALLVLILLIVCFIKRSRGGKYPVREKKDVPLGPED
PKEEDGSFDYSDEDNKPLQGSQTSLDGTIKQQESDDSLVDYGEGGEGQFNEDGSFIGQYT
VKKDKEETEGNESSEATSPVNAIYSLA
```

Uniprot – editing the sequence

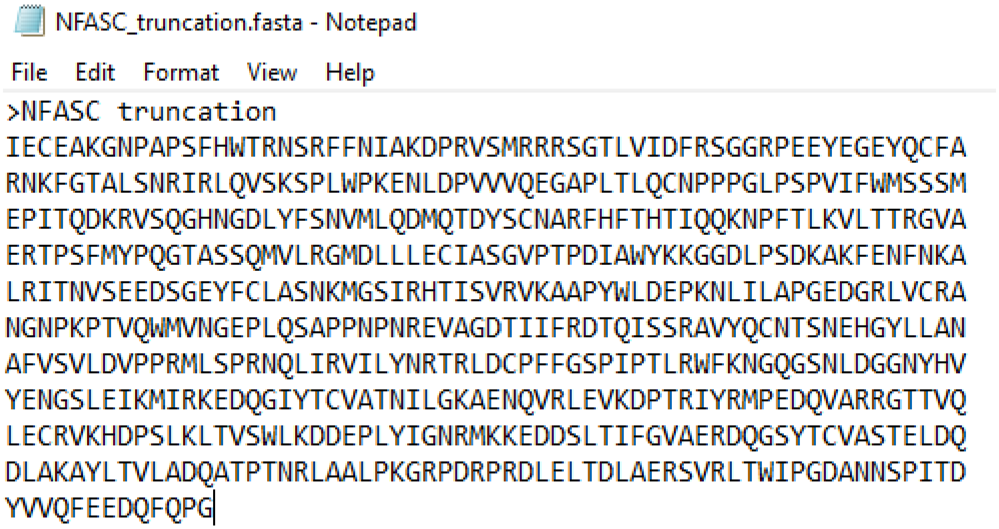
- But, what if I am only making one domain of this?



```
>sp|O94856|NFASC_HUMAN Neurofascin OS=Homo sapiens OX=9606 GN=NFASC PE=1 SV=4
MARQPPPPWVHAAFLLCLLSLGGAIEIPMDPSIQNELTQPPTITKQSAKDHIVDPRDNIL
IECEAKGNPAPSFHWTRNSRFFNIAKDPRVSMRRRSGTLVIDFRSGGRPEEYEGEYQCFA
RNKFGTALSNRIRLQVSKSPLWPKENLDPVVVQEGAPLTLQCNPPPGLPSPVIFWMSSSM
EPITQDKRVSQGHNGDLYFSNVMLQDMQTDYSCNARFHFTHTIQQKNPFTLKVLTTRGVA
ERTPSFMYPQGTASSQMVLRGMDLLLECIASGVPTPDIAWYKKGGDLPSDKAKFENFNKA
LRITNVSEEDSGEYFCLASNKMGSIRHTISVRVKAAPYWLDEPKNLILAPGEDGRLVCRA
NGNPKPTVQWMVNGEPLQSAPPNPNREVAGDTIIFRDTQISSRAVYQCNTSNEHGYLLAN
AFVSVLDVPPRMLSPRNQLIRVILYNRTRLDCPFFGSPIPTLRWFKNGQGSNLDGGNYHV
YENGSLEIKMIRKEDQGIYTCVATNILGKAENQVRLEVKDPTRIYRMPEDQVARRGTTVQ
LECRVKHDPSLKLTVSWLKDDEPLYIGNRMKKEDDSLTIFGVAERDQGSYTCVASTELDQ
DLAKAYLTVLADQATPTNRLAALPKGRPDRPRDLELTDLAERSVRLTWIPGDANNSPITD
YVVQFEEDQFQPGVWHDHSKYPGSVNSAVLRLSPYVNYQFRVIAINEVGSSHPSLPSERY
RTSGAPPESNPGDVKGEGTRKNNMEITWTPMNATSAFGPNLRYIVKWRRRETREAWNNVT
VWGSRYVVGQTPVYVPYEIRVQAENDFGKGPEPESVIGYSGEDYPRAAPTEVKVRVMNST
AISLQWNRVYSDTVQGQLREYRAYYWRESSLLKNLWVSQKRQQASFPGDRLRGVVSRLFP
YSNYKLEMVVVNGRGDGPRSETKEFTTPEGVPSAPRRFRVRQPNLETINLEWDHPEHPNG
IMIGYTLKYVAFNGTKVGKQIVENFSPNQTKFTVQRTDPVSRYRFTLSARTQVGSGEAVT
EESPAPPNEATPTAAPPTLPPTTVGATGAVSSTDATAIAATTEATTVPIIPTVAPTTIAT
TTTVATTTTTTAAATTTTESPPTTTSGTKIHESAPDEQSIWNVTVLPNSKWANITWKHNF
GPGTDFVVEYIDSNHTKKTVPVKAQAQPIQLTDLYPGMTYTLRVYSRDNEGISSTVITFM
TSTAYTNNQADIATQGWFIGLMCAIALLVLILLIVCFIKRSRGGKYPVREKKDVPLGPED
PKEEDGSFDYSDEDNKPLQGSQTSLDGTIKQQESDDSLVDYGEGGEGQFNEDGSFIGQYT
VKKDKEETEGNESSEATSPVNAIYSLA
```

Uniprot

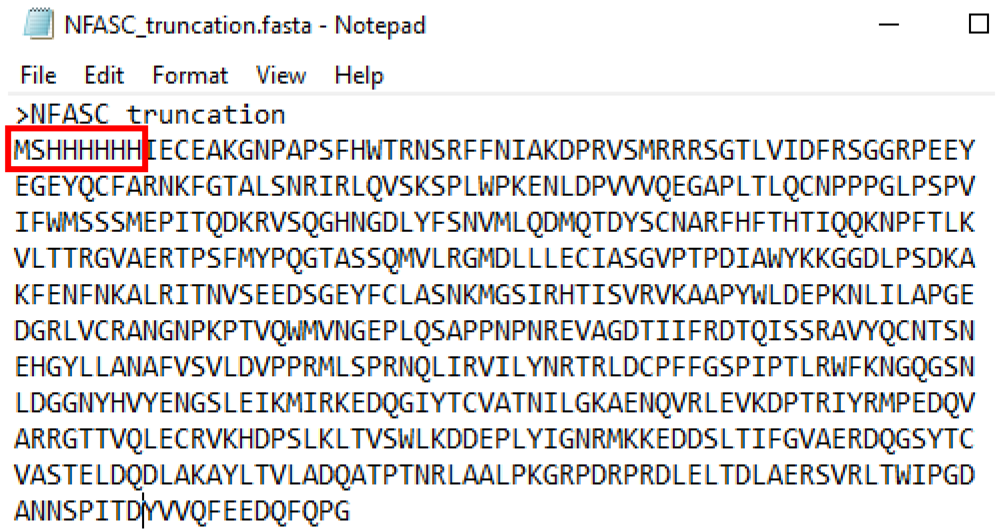
- You can manually edit this to contain your sequence of interest
- This can be done in any editing software, eg. NotePad



```
>NFASC truncation
IECEAKGNPAPSFHWTRNSRFFNIAKDPRVSMRRRSGTLVIDFRSGGRPEEYEGEYQCFA
RNKFGTALSNRIRLQVSKSPLWPKENLDPVVVQEGAPLTLQCNPPPGLPSPVIFWMSSSM
EPITQDKRVSQGHNGDLYFSNVMLQDMQTDYSCNARFHFTHTIQQKNPFTLKVLTTRGVA
ERTPSFMYPQGTASSQMVLRGMDLLLECIASGVPTPDIAWYKKGGDLPSDKAKFENFNKA
LRITNVSEEDSGEYFCLASNKMGSIRHTISVRVKAAPYWLDEPKNLILAPGEDGRLVCRA
NGNPKPTVQWMVNGEPLQSAPPNPNREVAGDTIIFRDTQISSRAVYQCNTSNEHGYLLAN
AFVSVLDVPPRMLSPRNQLIRVILYNRTRLDCPFFGSPIPTLRWFKNGQGSNLDGGNYHV
YENGSLEIKMIRKEDQGIYTCVATNILGKAENQVRLEVKDPTRIYRMPEDQVARRGTTVQ
LECRVKHDPSLKLTVSWLKDDEPLYIGNRMKKEDDSLTIFGVAERDQGSYTCVASTELDQ
DLAKAYLTVLADQATPTNRLAALPKGRPDRPRDLELTDLAERSVRLTWIPGDANNSPITD
YVVQFEEDQFQPG
```

Uniprot

- You can manually edit this to contain your sequence of interest
- This can be done in any editing software, eg. NotePad

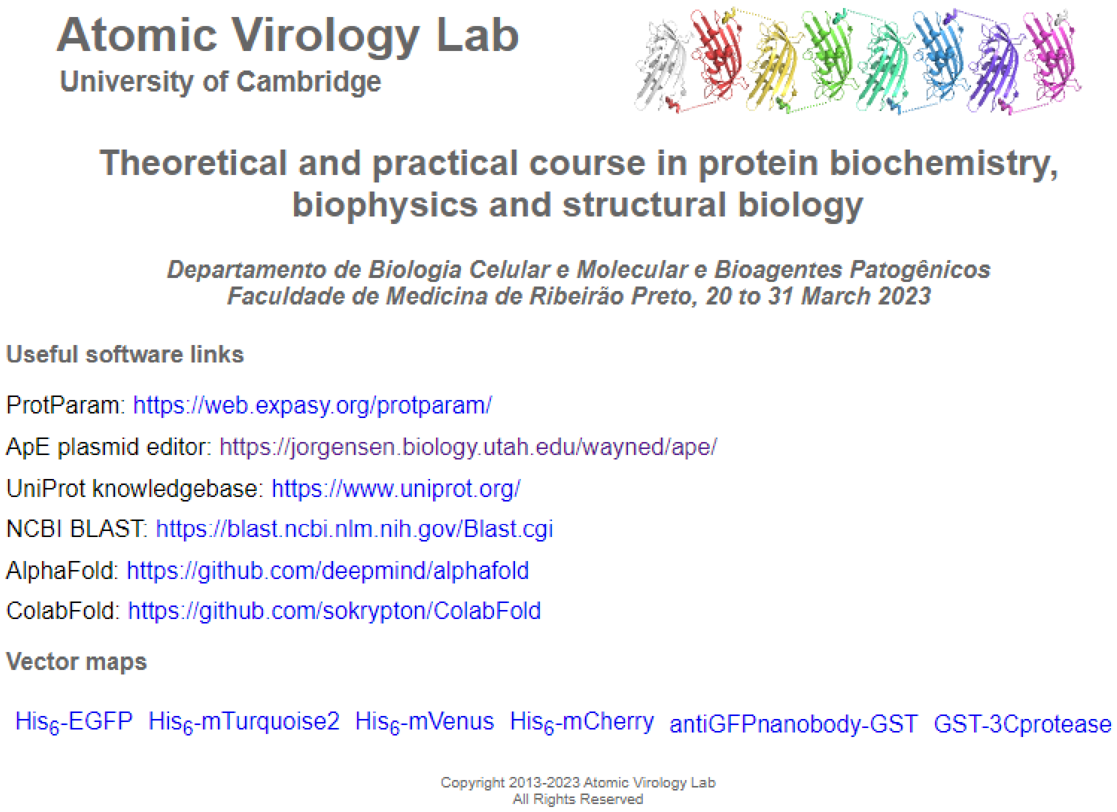


```
>NFASC truncation
MSHHHHHHIECEAKGNPAPSFHWTRNSRFFNIAKDPRVSMRRRSGTLVIDFRSGGRPEEY
EGEYQCFARNKFGTALSNRIRLQVSKSPLWPKENLDPVVVQEGAPLTLQCNPPPGLPSPV
IFWMSSSMEPITQDKRVSQGHNGDLYFSNVMLQDMQTDYSCNARFHFTHTIQQKNPFTLK
VLTTRGVAERTPSFMYPQGTASSQMVLRGMDLLLECIASGVPTPDIAWYKKGGDLPSDKA
KFENFNKALRITNVSEEDSGEYFCLASNKMGSIRHTISVRVKAAPYWLDEPKNLILAPGE
DGRLVCRANGNPKPTVQWMVNGEPLQSAPPNPNREVAGDTIIFRDTQISSRAVYQCNTSN
EHGYLLANAFVSVLDVPPRMLSPRNQLIRVILYNRTRLDCPFFGSPIPTLRWFKNGQGSN
LDGGNYHVYENGSLEIKMIRKEDQGIYTCVATNILGKAENQVRLEVKDPTRIYRMPEDQV
ARRGTTVQLECRVKHDPSLKLTVSWLKDDEPLYIGNRMKKEDDSLTIFGVAERDQGSYTC
VASTELDQDLAKAYLTVLADQATPTNRLAALPKGRPDRPRDLELTDLAERSVRLTWIPGD
ANNSPITDYVVQFEEDQFQPG
```

- You can also add the sequence of tags such as His or GST-tags

Or you can get the sequence from a plasmid map

- For this course, there is a link to the software and plasmids files we're using: http://www.atomicvirology.path.cam.ac.uk/brazil



Atomic Virology Lab

University of Cambridge

Theoretical and practical course in protein biochemistry, biophysics and structural biology

Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos
Faculdade de Medicina de Ribeirão Preto, 20 to 31 March 2023

Useful software links

ProtParam: https://web.expasy.org/protparam/

ApE plasmid editor: https://jorgensen.biology.utah.edu/wayned/ape/

UniProt knowledgebase: https://www.uniprot.org/

NCBI BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi

AlphaFold: https://github.com/deepmind/alphafold

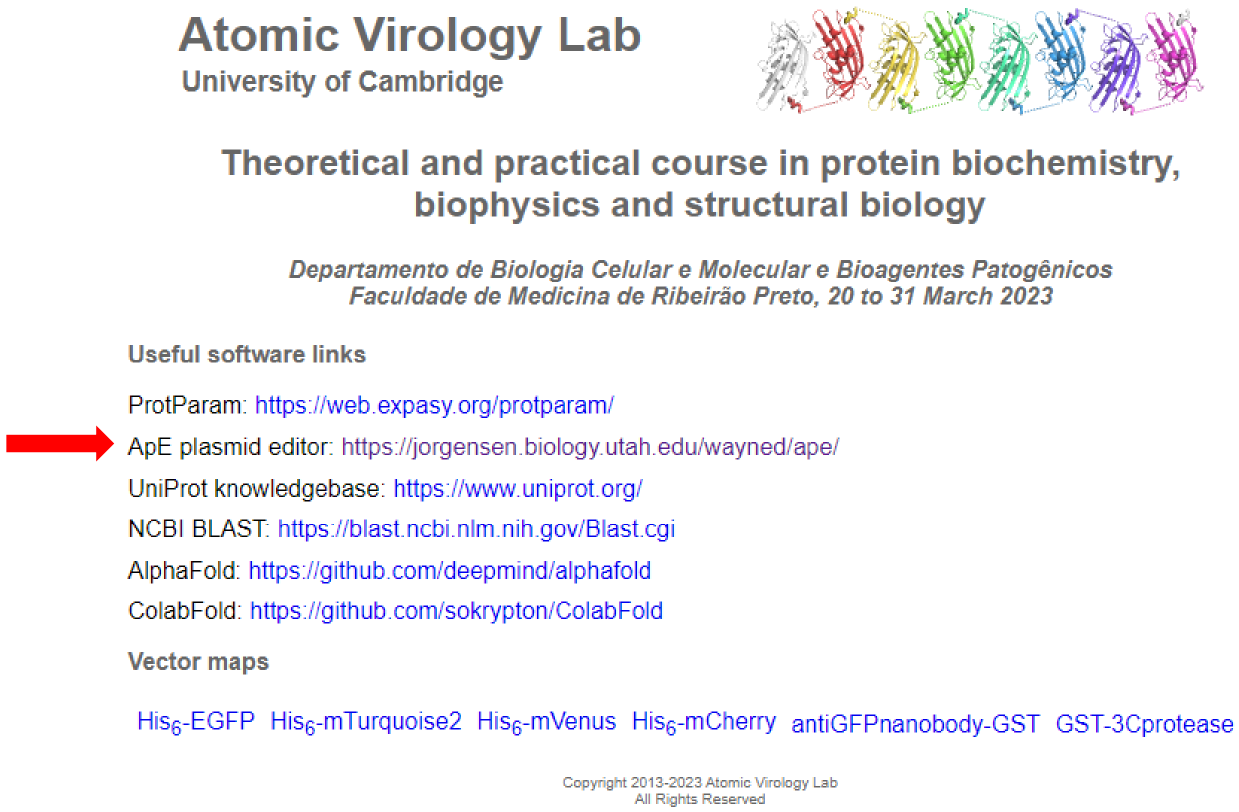
ColabFold: https://github.com/sokrypton/ColabFold

Vector maps

His$_6$-EGFP His$_6$-mTurquoise2 His$_6$-mVenus His$_6$-mCherry antiGFPnanobody-GST GST-3Cprotease

Or you can get the sequence from a plasmid map

- Use ApE to open the plasmid file and extract the sequence data



Atomic Virology Lab
University of Cambridge

Theoretical and practical course in protein biochemistry, biophysics and structural biology

Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos
Faculdade de Medicina de Ribeirão Preto, 20 to 31 March 2023

Useful software links

ProtParam: https://web.expasy.org/protparam/

ApE plasmid editor: https://jorgensen.biology.utah.edu/wayned/ape/

UniProt knowledgebase: https://www.uniprot.org/

NCBI BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi

AlphaFold: https://github.com/deepmind/alphafold

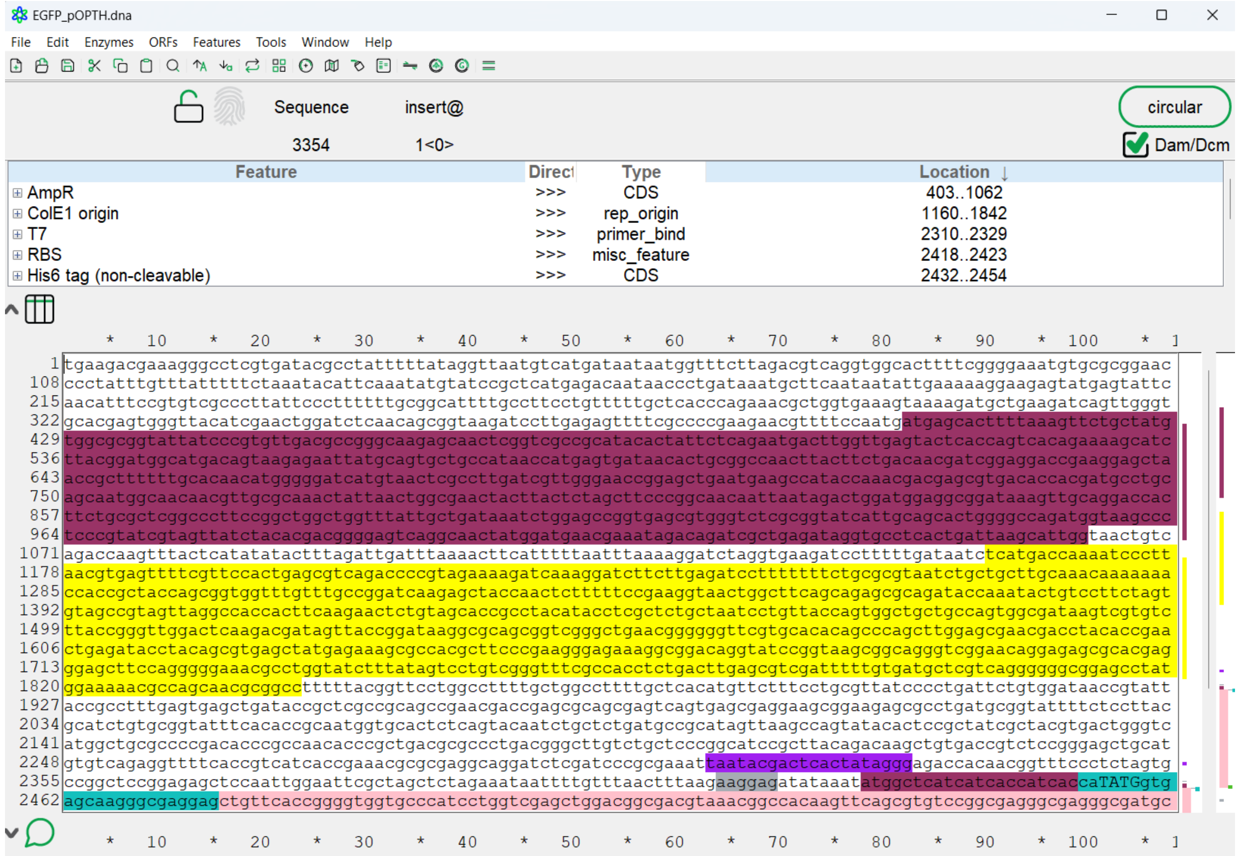
ColabFold: https://github.com/sokrypton/ColabFold

Vector maps

His$_6$-EGFP His$_6$-mTurquoise2 His$_6$-mVenus His$_6$-mCherry antiGFPnanobody-GST GST-3Cprotease

ApE – A plasmid Editor

- Open your vector map – different features are highlighted



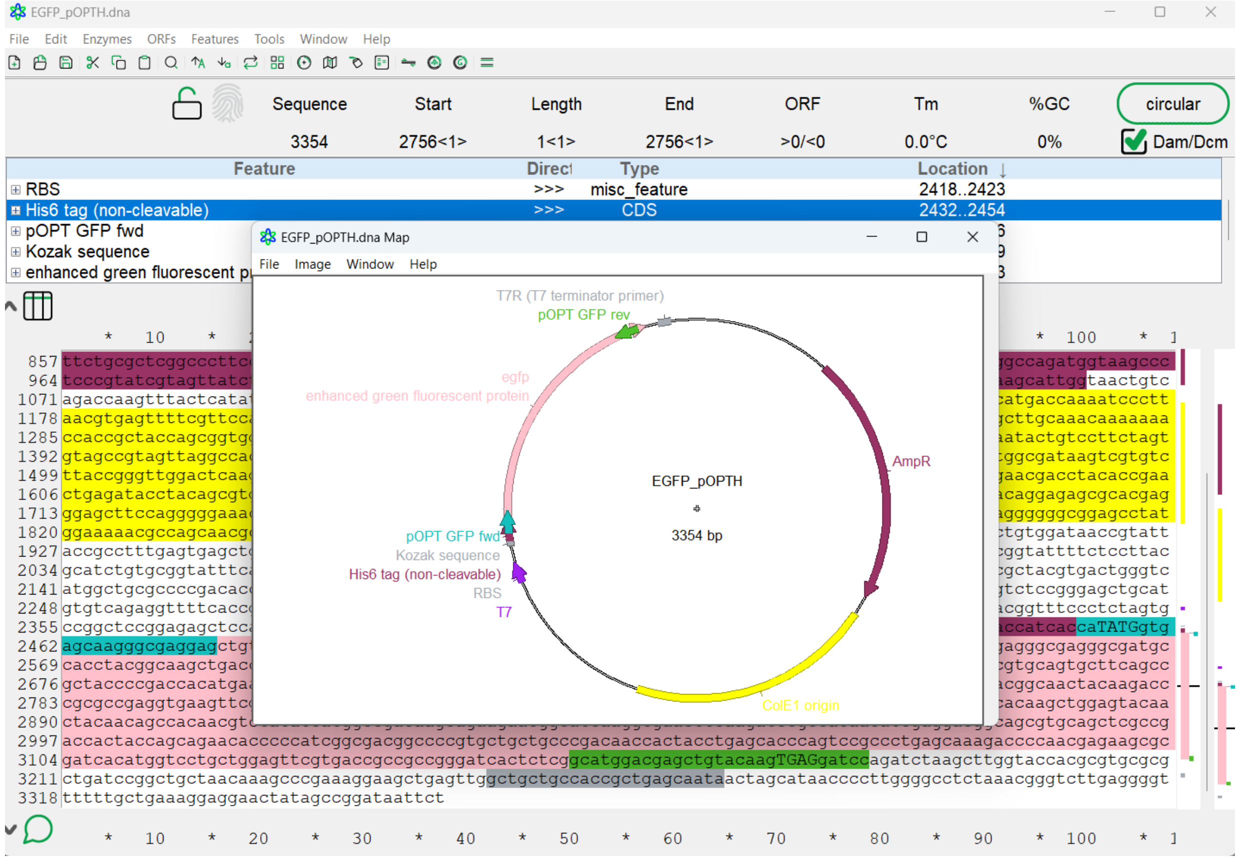
Sequence	insert@
3354	1<0>

Feature	Direct	Type	Location ↓
⊞ AmpR	>>>	CDS	403..1062
⊞ ColE1 origin	>>>	rep_origin	1160..1842
⊞ T7	>>>	primer_bind	2310..2329
⊞ RBS	>>>	misc_feature	2418..2423
⊞ His6 tag (non-cleavable)	>>>	CDS	2432..2454

```
        *   10     *   20     *   30     *   40     *   50     *   60     *   70     *   80     *   90     *  100     *  1
   1 tgaagacgaaagggcctcgtgatacgcctatttttataggttaatgtcatgataataatggtttcttagacgtcaggtggcacttttcggggaaatgtgcgcggaac
 108 ccctatttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatgagtattc
 215 aacatttccgtgtcgcccttattcccttttttgcggcattttgccttcctgtttttgctcacccagaaacgctggtgaaagtaaaagatgctgaagatcagttgggt
 322 gcacgagtgggttacatcgaactggatctcaacagcggtaagatccttgagagttttcgccccgaagaacgttttccaatgatgagcacttttaaagttctgctatg
 429 tggcgcggtattatcccgtgttgacgccgggcaagagcaactcggtcgccgcatacactattctcagaatgacttggttgagtactcaccagtcacagaaaagcatc
 536 ttacggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacactgcggccaacttacttctgacaacgatcggaggaccgaaggagcta
 643 accgcttttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaaccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgc
 750 agcaatggcaacaacgttgcgcaaactattaactggcgaactacttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccac
 857 ttctgcgctcggcccttccggctggctggtttattgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccc
 964 tcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtc
1071 agaccaagtttactcatatatactttagattgatttaaaacttcattttaatttaaaaggatctaggtgaagatcctttttgataatctcatgaccaaaatccctt
1178 aacgtgagttttcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaa
1285 ccaccgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagt
1392 gtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtc
1499 ttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaa
1606 ctgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgag
1713 ggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctat
1820 ggaaaaacgccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtatt
1927 accgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcggtattttctccttac
2034 gcatctgtgcggtatttcacaccgcaatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgctatcgctacgtgactgggtc
2141 atggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcat
2248 gtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcaggatctcgatcccgcgaaattaatacgactcactatagggagaccacaacggtttccctctagtg
2355 ccggctccggagagctccaattggaattcgctagctctagaaataattttgtttaactttaagaaggagatataaatatggctcatcatcaccatcaccaTATGgtg
2462 agcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgc
        *   10     *   20     *   30     *   40     *   50     *   60     *   70     *   80     *   90     *  100     *  1
```

ApE – A plasmid Editor

- You can also view this as a graphic map



EGFP_pOPTH.dna

Sequence	Start	Length	End	ORF	Tm	%GC	circular
3354	2756<1>	1<1>	2756<1>	>0/<0	0.0°C	0%	☑ Dam/Dcm

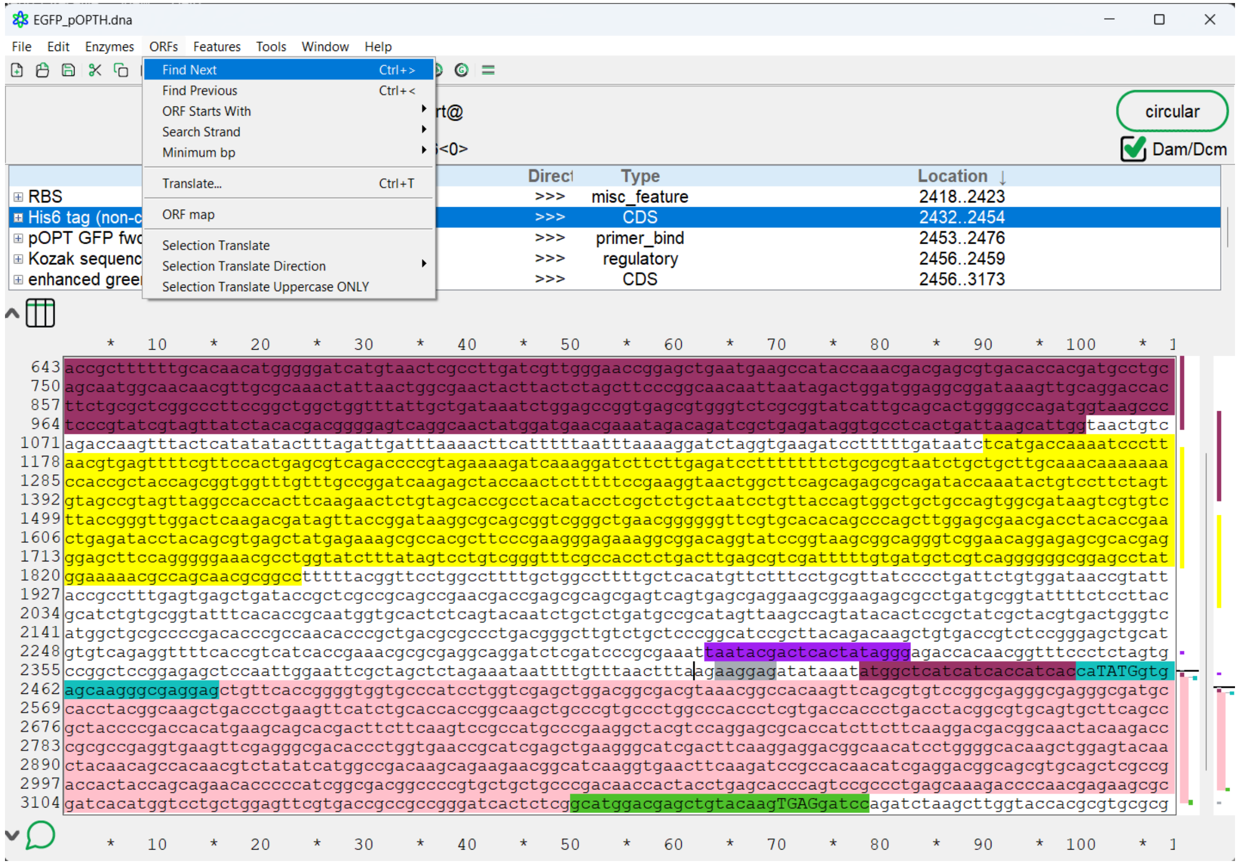
Feature	Directi	Type	Location ↓
⊞ RBS	>>>	misc_feature	2418..2423
⊞ His6 tag (non-cleavable)	>>>	CDS	2432..2454
⊞ pOPT GFP fwd			6
⊞ Kozak sequence			9
⊞ enhanced green fluorescent p			3

```
       *   10    *                                    *  100     *  1
 857 ttctgcgctcggccctt                          gccagatggtaagccc
 964 tcccgtatcgtagttatc                         agcattggtaactgtc
1071 agaccaagtttactcata                         catgaccaaaatccctt
1178 aacgtgagttttcgttcc                         cttgcaaacaaaaaaa
1285 ccaccgctaccagcggtg                          atactgtccttctagt
1392 gtagccgtagttaggcca                          ggcgataagtcgtgtc
1499 ttaccgggttggactcaa                          aacgacctacaccgaa
1606 ctgagatacctacagcgt                          caggagagcgcacgag
1713 ggagcttccaggggaaa                          gggggcggagcctat
1820 ggaaaaacgccagcaacg                         tgtggataaccgtatt
1927 accgcctttgagtgagct                         ggtattttctccttac
2034 gcatctgtgcggtatttc                         gctacgtgactgggtc
2141 atggctgcgccccgacac                          tctccgggagctgcat
2248 gtgtcagaggttttcacc                          cggtttccctctagtg
2355 ccggctccggagagctcc                          ccatcaccaTATGgtg
2462 agcaagggcgaggagctg                          agggcgagggcgatgc
2569 cacctacggcaagctgac                          gtgcagtgcttcagcc
2676 gctaccccgaccacatga                          cggcaactacaagacc
2783 cgcgccgaggtgaagttc                          acaagctggagtacaa
2890 ctacaacagccacaacgt                          agcgtgcagctcgccg
2997 accactaccagcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgc
3104 gatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagTGAGgatccagatctaagcttggtaccacgcgtgcgcg
3211 ctgatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcataaccccttggggcctctaaacgggtcttgaggggt
3318 tttttgctgaaaggaggaactatagccggataattct
       *   10    *   20    *   30    *   40    *   50    *   60    *   70    *   80    *   90    *  100     *  1
```

EGFP_pOPTH.dna Map

ApE – A plasmid Editor

- Identify the relevant open reading frame (ORF)



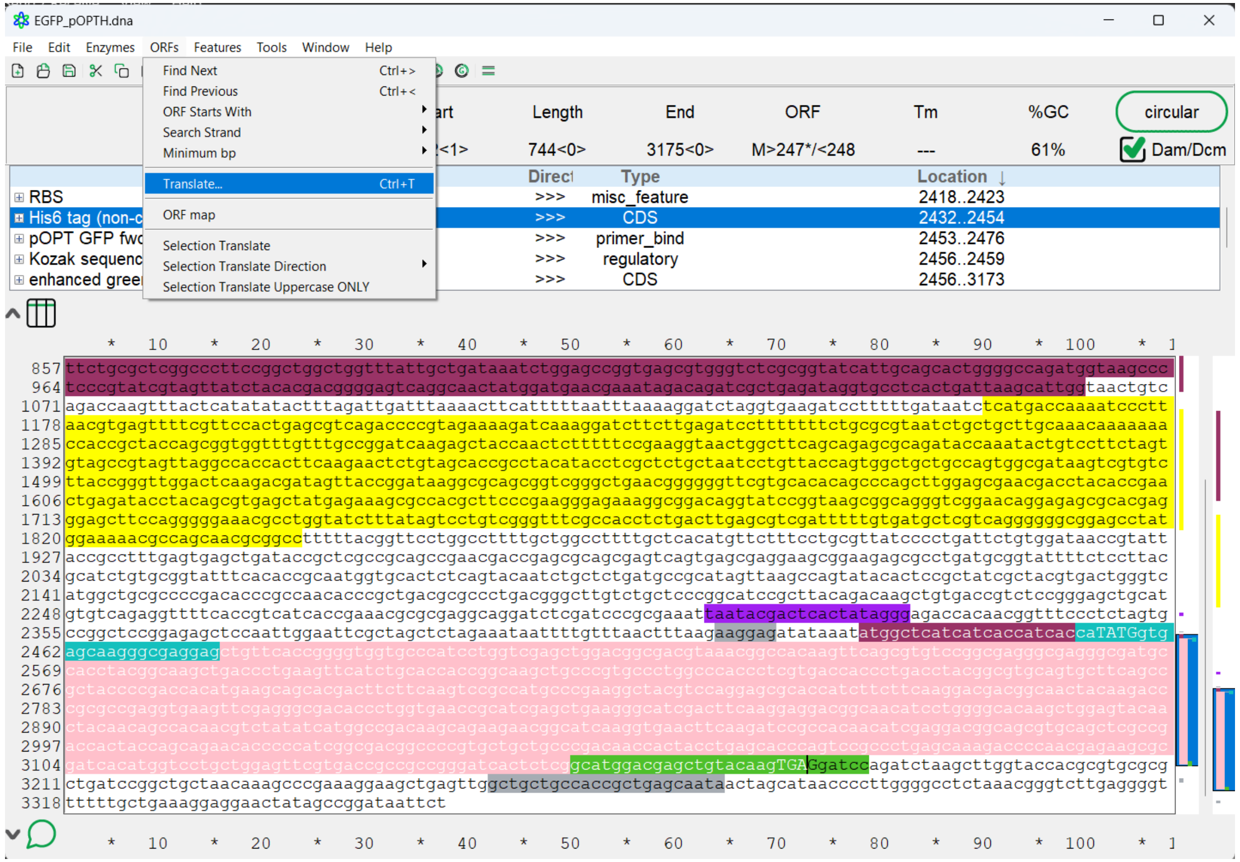
	Direct	Type	Location ↓
RBS	>>>	misc_feature	2418..2423
His6 tag (non-c	>>>	CDS	2432..2454
pOPT GFP fwd	>>>	primer_bind	2453..2476
Kozak sequenc	>>>	regulatory	2456..2459
enhanced greer	>>>	CDS	2456..3173

```
       *   10    *   20    *   30    *   40    *   50    *   60    *   70    *   80    *   90    *  100    *  1
 643 accgcttttttgcacaacatgggggatcatgtaactcgccttgatcgttgggaaccggagctgaatgaagccataccaaacgacgagcgtgacaccacgatgcctgc
 750 agcaatggcaacaacgttgcgcaaactattaactggcgaactacttactctagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggaccac
 857 ttctgcgctcggcccttccggctggctggtttattgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccc
 964 tcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtc
1071 agaccaagtttactcatatatactttagattgatttaaaacttcatttttaatttaaaaggatctaggtgaagatcctttttgataatctcatgaccaaaatccctt
1178 aacgtgagttttcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaa
1285 ccaccgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagt
1392 gtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtc
1499 ttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaa
1606 ctgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgag
1713 ggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctat
1820 ggaaaaacgccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtatt
1927 accgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcggtattttctccttac
2034 gcatctgtgcggtatttcacaccgcaatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgctatcgctacgtgactgggtc
2141 atggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcat
2248 gtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcaggatctcgatcccgcgaaattaatacgactcactatagggagaccacaacggtttccctctagtg
2355 ccggctccggagagctccaattggaattcgctagctctagaaataattttgtttaactttaagaaggagatataaatatggctcatcatcaccatcaccaTATGgtg
2462 agcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgc
2569 cacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagcc
2676 gctaccccgaccacatgaagcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacc
2783 cgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaa
2890 ctacaacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccg
2997 accactaccagcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgc
3104 gatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagTGAGgatccagatctaagcttggtaccacgcgtgcgcg
       *   10    *   20    *   30    *   40    *   50    *   60    *   70    *   80    *   90    *  100    *  1
```

UNIVERSITY OF CAMBRIDGE

ApE – A plasmid Editor

- Translate the ORF



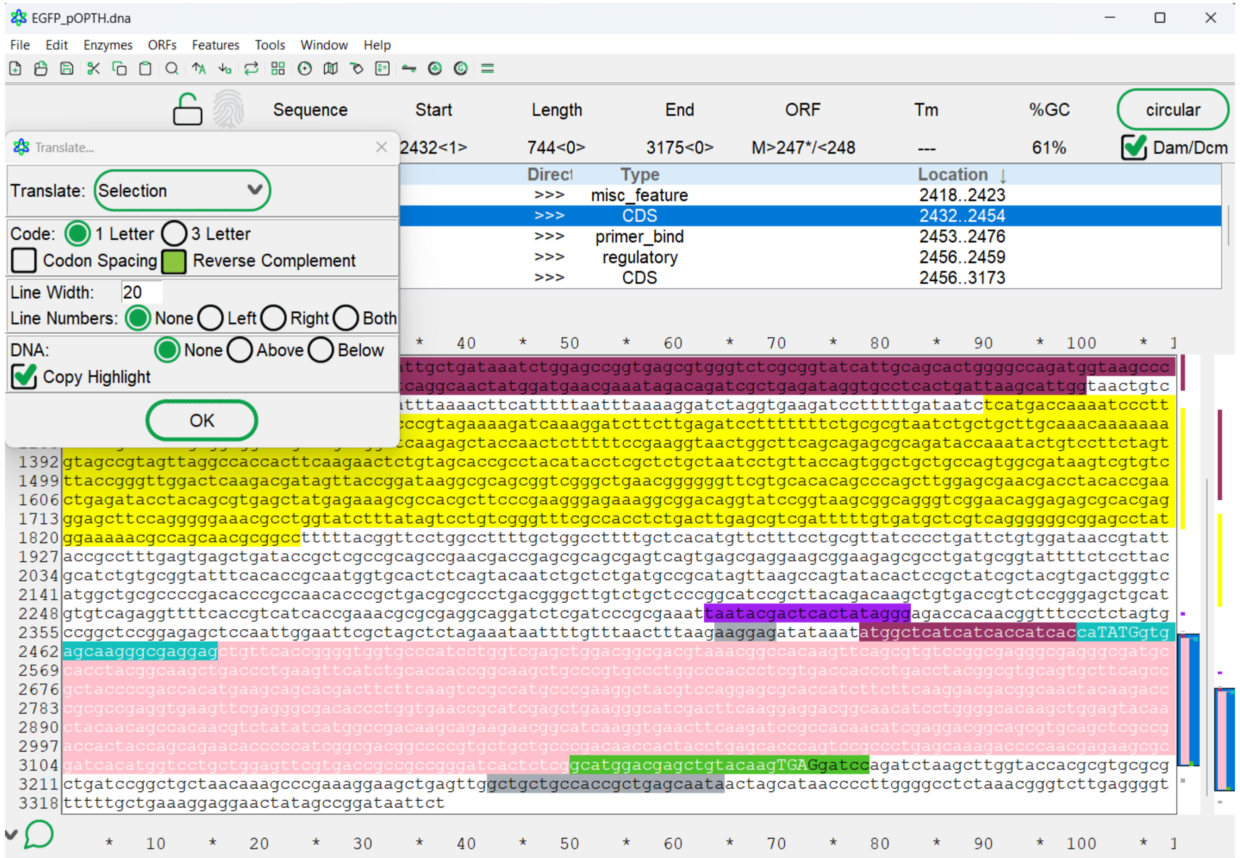
art	Length	End	ORF	Tm	%GC
<1>	744<0>	3175<0>	M>247*/<248	---	61%

	Direct	Type	Location ↓
RBS	>>>	misc_feature	2418..2423
His6 tag (non-c	>>>	CDS	2432..2454
pOPT GFP fwd	>>>	primer_bind	2453..2476
Kozak sequenc	>>>	regulatory	2456..2459
enhanced gree	>>>	CDS	2456..3173

```
      *  10   *  20   *  30   *  40   *  50   *  60   *  70   *  80   *  90   * 100   * 1
 857 ttctgcgctcggcccttccggctggctggtttattgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccc
 964 tcccgtatcgtagttatctacacgacggggagtcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtc
1071 agaccaagtttactcatatatactttagattgatttaaaacttcatttttaatttaaaaggatctaggtgaagatcctttttgataatctcatgaccaaaatccctt
1178 aacgtgagttttcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaa
1285 ccaccgctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagt
1392 gtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtc
1499 ttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaa
1606 ctgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgag
1713 ggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctat
1820 ggaaaaacgccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtatt
1927 accgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcggtattttctccttac
2034 gcatctgtgcggtatttcacaccgcaatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgctatcgctacgtgactgggtc
2141 atggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcat
2248 gtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcaggatctcgatcccgcgaaattaatacgactcactatagggagaccacaacggtttccctctagtg
2355 ccggctccggagagctccaattggaattcgctagctctagaaataattttgtttaactttaagaaggagatataaatatggctcatcatcaccatcaccaTATGgtg
2462 agcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgc
2569 cacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagcc
2676 gctaccccgaccacatgaagcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacc
2783 cgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaa
2890 ctacaacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccg
2997 accactaccagcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgc
3104 gatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagTGAGgatccagatctaagcttggtaccacgcgtgcgcg
3211 ctgatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcataaccccttggggcctctaaacgggtcttgaggggt
3318 tttttgctgaaaggaggaactatagccggataattct
      *  10   *  20   *  30   *  40   *  50   *  60   *  70   *  80   *  90   * 100   * 1
```

ApE – A plasmid Editor

- Translate as 1-letter code



Sequence	Start	Length	End	ORF	Tm	%GC	circular
	2432<1>	744<0>	3175<0>	M>247*/<248	---	61%	☑ Dam/Dcm

Translate: Selection

Code: ● 1 Letter ○ 3 Letter

□ Codon Spacing ■ Reverse Complement

Line Width: 20

Line Numbers: ● None ○ Left ○ Right ○ Both

DNA: ● None ○ Above ○ Below

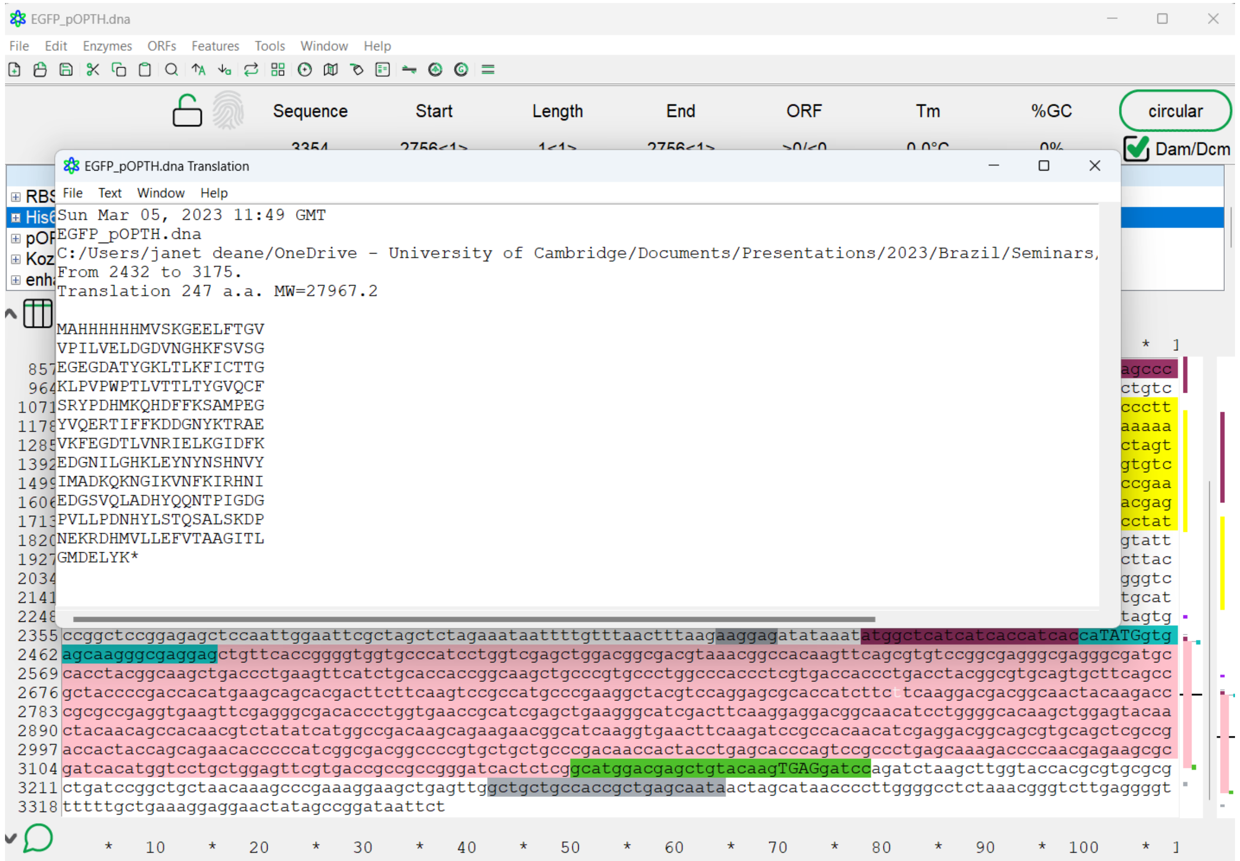
☑ Copy Highlight

Direct	Type	Location ↓
>>>	misc_feature	2418..2423
>>>	CDS	2432..2454
>>>	primer_bind	2453..2476
>>>	regulatory	2456..2459
>>>	CDS	2456..3173

```
     *   40    *   50    *   60    *   70    *   80    *   90    *  100    *  1
     attgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgcagcactggggccagatggtaagccc
     tcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgattaagcattggtaactgtc
     atttaaaacttcatttttaatttaaaaggatctaggtgaagatcctttttgataatctcatgaccaaaatccctt
     cccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaa
     ccaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagt
1392 gtagccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtc
1499 ttaccgggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttcgtgcacacagcccagcttggagcgaacgacctacaccgaa
1606 ctgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacaggagagcgcacgag
1713 ggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggggcggagcctat
1820 ggaaaaacgccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataaccgtatt
1927 accgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcggtattttctccttac
2034 gcatctgtgcggtatttcacaccgcaatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgctatcgctacgtgactgggtc
2141 atggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcat
2248 gtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcaggatctcgatcccgcgaaattaatacgactcactatagggagaccacaacggtttccctctagtg
2355 ccggctccggagagctccaattggaattcgctagctctagaaataattttgtttaactttaagaaggagatataaatatggctcatcatcaccatcaccaTATGgtg
2462 agcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgc
2569 cacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagcc
2676 gctaccccgaccacatgaagcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacc
2783 cgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaa
2890 ctacaacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccg
2997 accactaccagcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgc
3104 gatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagTGAGgatccagatctaagcttggtaccacgcgtgcgcg
3211 ctgatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcataaccccttggggcctctaaacgggtcttgaggggt
3318 ttttgctgaaaggaggaactatagccggataattct
     *   10    *   20    *   30    *   40    *   50    *   60    *   70    *   80    *   90    *  100    *  1
```

ApE – A plasmid Editor

- Now you have the exact sequence of your construct including tags etc



```
Sun Mar 05, 2023 11:49 GMT
EGFP_pOPTH.dna
C:/Users/janet deane/OneDrive - University of Cambridge/Documents/Presentations/2023/Brazil/Seminars,
From 2432 to 3175.
Translation 247 a.a. MW=27967.2

MAHHHHHHMVSKGEELFTGV
VPILVELDGDVNGHKFSVSG
EGEGDATYGKLTLKFICTTG
KLPVPWPTLVTTLTYGVQCF
SRYPDHMKQHDFFKSAMPEG
YVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFK
EDGNILGHKLEYNYNSHNVY
IMADKQKNGIKVNFKIRHNI
EDGSVQLADHYQQNTPIGDG
PVLLPDNHYLSTQSALSKDP
NEKRDHMVLLEFVTAAGITL
GMDELYK*
```

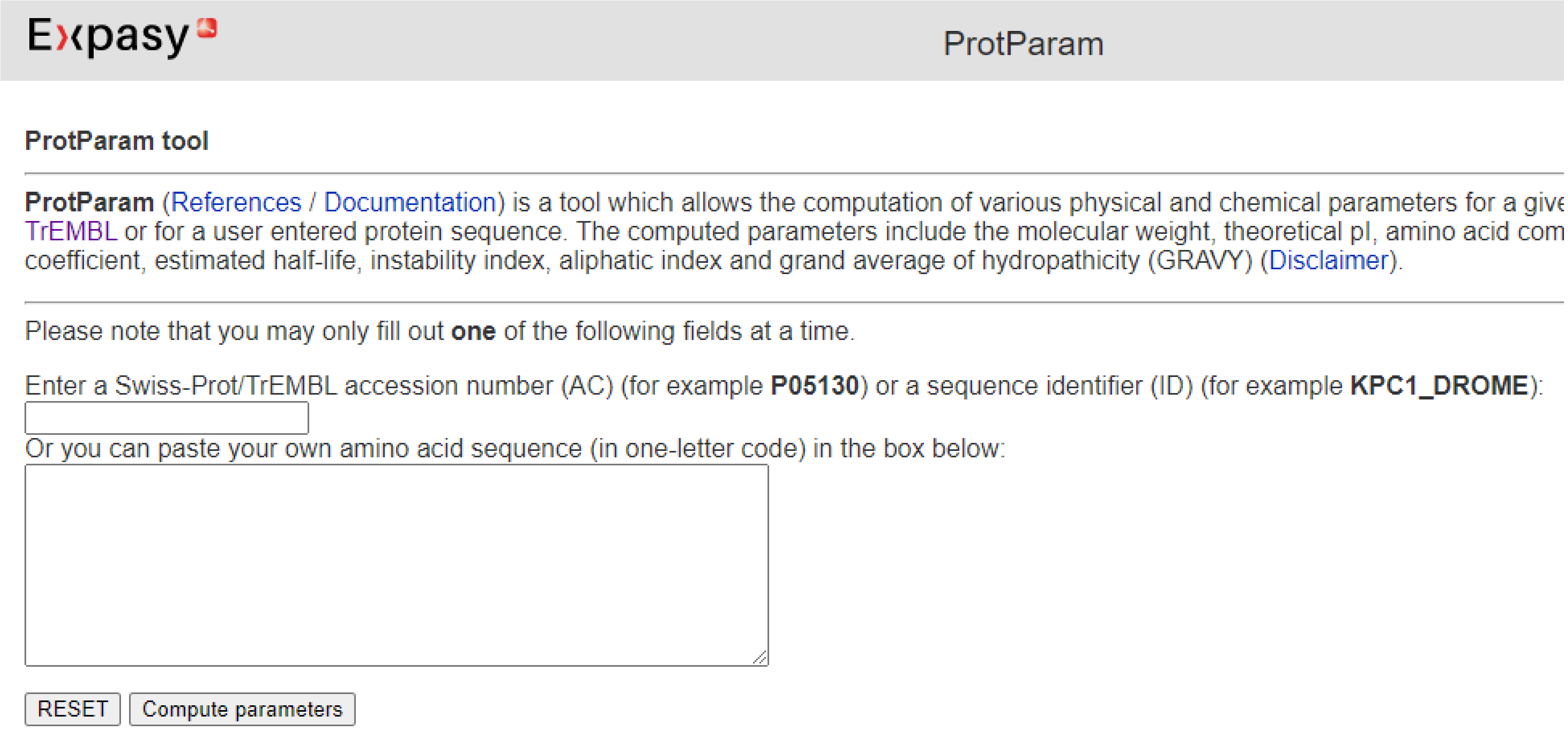
```
2355 ccggctccggagagctccaattggaattcgctagctctagaaataattttgtttaactttaagaaggagatataaatatggctcatcatcaccatcaccaTATGgtg
2462 agcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgc
2569 cacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagcc
2676 gctaccccgaccacatgaagcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacc
2783 cgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaa
2890 ctacaacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccg
2997 accactaccagcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccccaacgagaagcgc
3104 gatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagTGAGgatccagatctaagcttggtaccacgcgtgcgcg
3211 ctgatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcataaccccttggggcctctaaacgggtcttgaggggt
3318 ttttgctgaaaggaggaactatagccggataattct
          *   10    *   20    *   30    *   40    *   50    *   60    *   70    *   80    *   90    *  100    *  1
```

Second useful resource - ProtParam

- Protparam uses a sequence to calculate:
  - molecular mass
  - isoelectric point
  - extinction coefficients

https://web.expasy.org/protparam/

Second useful resource - ProtParam



ProtParam tool

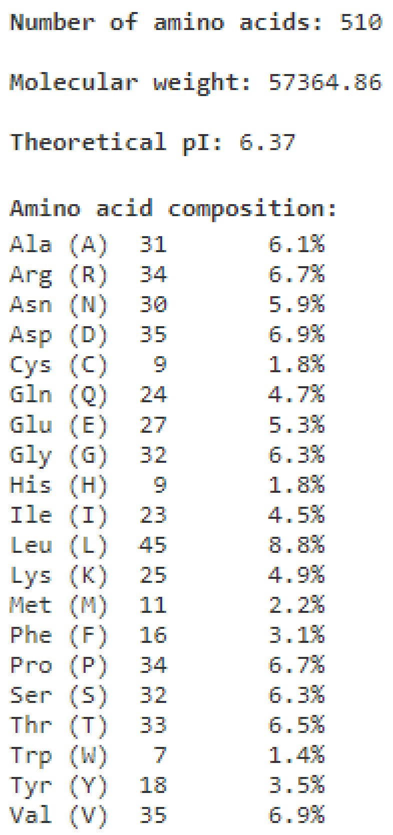
ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a give TrEMBL or for a user entered protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid com coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

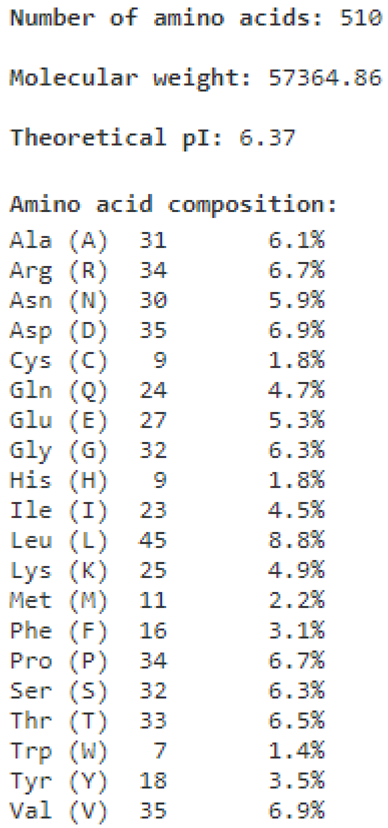
Or you can paste your own amino acid sequence (in one-letter code) in the box below:

Second useful resource - ProtParam



```
Number of amino acids: 510

Molecular weight: 57364.86

Theoretical pI: 6.37

Amino acid composition:
Ala (A)  31        6.1%
Arg (R)  34        6.7%
Asn (N)  30        5.9%
Asp (D)  35        6.9%
Cys (C)   9        1.8%
Gln (Q)  24        4.7%
Glu (E)  27        5.3%
Gly (G)  32        6.3%
His (H)   9        1.8%
Ile (I)  23        4.5%
Leu (L)  45        8.8%
Lys (K)  25        4.9%
Met (M)  11        2.2%
Phe (F)  16        3.1%
Pro (P)  34        6.7%
Ser (S)  32        6.3%
Thr (T)  33        6.5%
Trp (W)   7        1.4%
Tyr (Y)  18        3.5%
Val (V)  35        6.9%
```

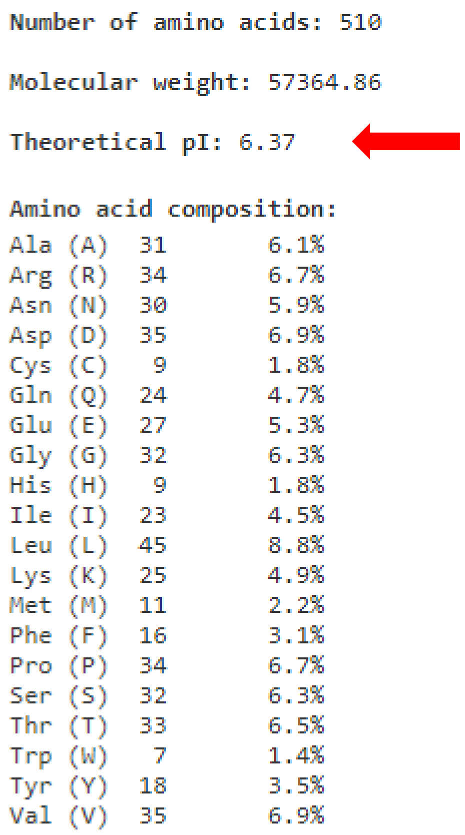
Second useful resource - ProtParam



```
Number of amino acids: 510

Molecular weight: 57364.86

Theoretical pI: 6.37

Amino acid composition:
Ala (A)  31        6.1%
Arg (R)  34        6.7%
Asn (N)  30        5.9%
Asp (D)  35        6.9%
Cys (C)   9        1.8%
Gln (Q)  24        4.7%
Glu (E)  27        5.3%
Gly (G)  32        6.3%
His (H)   9        1.8%
Ile (I)  23        4.5%
Leu (L)  45        8.8%
Lys (K)  25        4.9%
Met (M)  11        2.2%
Phe (F)  16        3.1%
Pro (P)  34        6.7%
Ser (S)  32        6.3%
Thr (T)  33        6.5%
Trp (W)   7        1.4%
Tyr (Y)  18        3.5%
Val (V)  35        6.9%
```

Mass in daltons, 57.4 kDa

Second useful resource - ProtParam



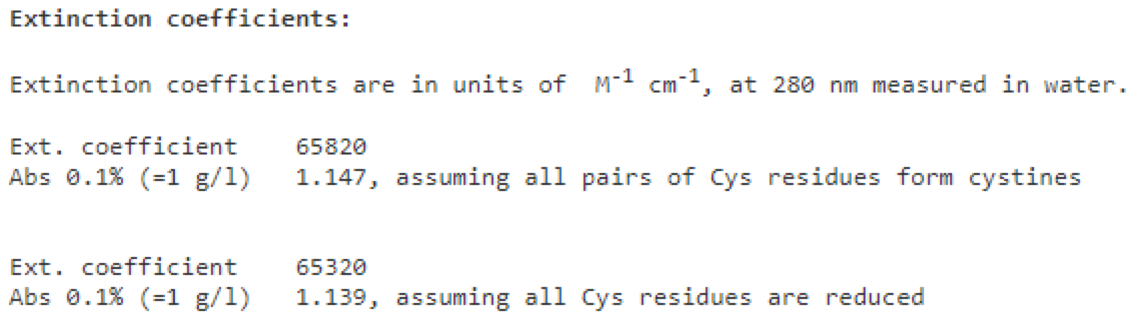
```
Number of amino acids: 510

Molecular weight: 57364.86

Theoretical pI: 6.37

Amino acid composition:
Ala (A)  31        6.1%
Arg (R)  34        6.7%
Asn (N)  30        5.9%
Asp (D)  35        6.9%
Cys (C)   9        1.8%
Gln (Q)  24        4.7%
Glu (E)  27        5.3%
Gly (G)  32        6.3%
His (H)   9        1.8%
Ile (I)  23        4.5%
Leu (L)  45        8.8%
Lys (K)  25        4.9%
Met (M)  11        2.2%
Phe (F)  16        3.1%
Pro (P)  34        6.7%
Ser (S)  32        6.3%
Thr (T)  33        6.5%
Trp (W)   7        1.4%
Tyr (Y)  18        3.5%
Val (V)  35        6.9%
```

pI is the pH where your protein has no charge

Your protein can be unstable at its pI

Keep your purification buffers at least 1 pH unit from the pI

Extinction coefficients for protein concentration

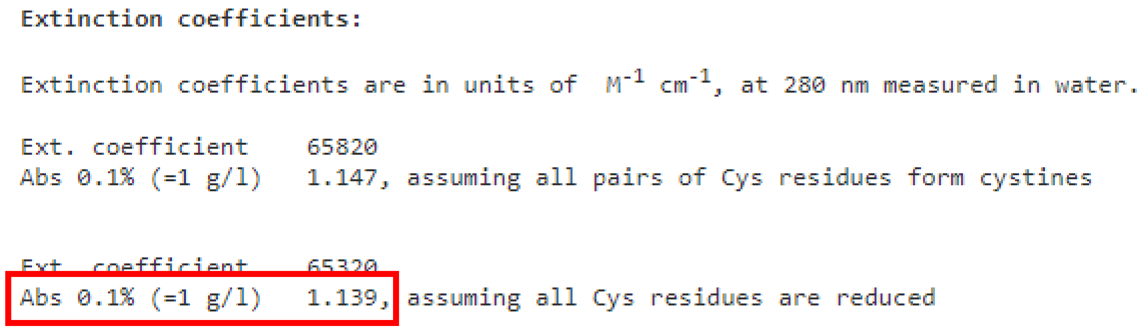


Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

```
Ext. coefficient    65820
Abs 0.1% (=1 g/l)   1.147, assuming all pairs of Cys residues form cystines


Ext. coefficient    65320
Abs 0.1% (=1 g/l)   1.139, assuming all Cys residues are reduced
```

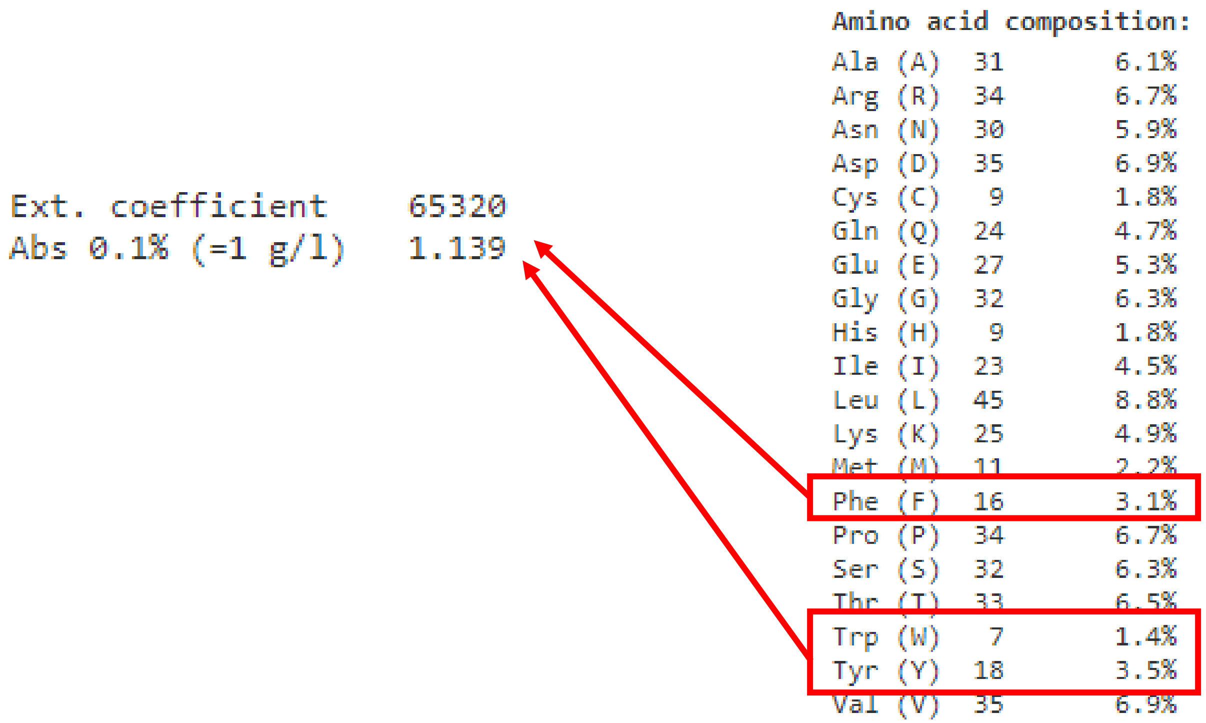
Extinction coefficients for protein concentration



```
Extinction coefficients:

Extinction coefficients are in units of  M-1 cm-1, at 280 nm measured in water.

Ext. coefficient    65820
Abs 0.1% (=1 g/l)   1.147, assuming all pairs of Cys residues form cystines


Ext. coefficient    65320
Abs 0.1% (=1 g/l)   1.139, assuming all Cys residues are reduced
```

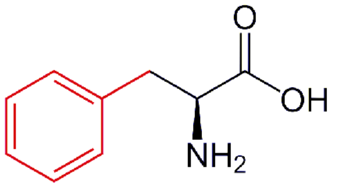
$E_{0.1\%}$ in units of $(mg/mL)^{-1}$

What determines the E value?

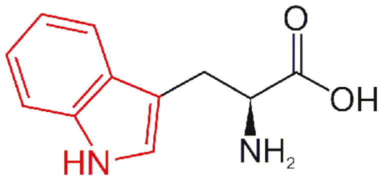


```
Ext. coefficient      65320
Abs 0.1% (=1 g/l)     1.139
```

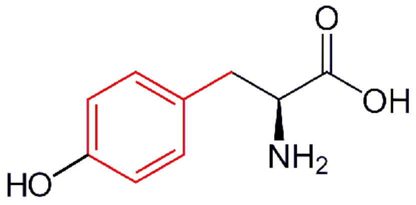
```
Amino acid composition:
Ala (A)  31         6.1%
Arg (R)  34         6.7%
Asn (N)  30         5.9%
Asp (D)  35         6.9%
Cys (C)   9         1.8%
Gln (Q)  24         4.7%
Glu (E)  27         5.3%
Gly (G)  32         6.3%
His (H)   9         1.8%
Ile (I)  23         4.5%
Leu (L)  45         8.8%
Lys (K)  25         4.9%
Met (M)  11         2.2%
Phe (F)  16         3.1%
Pro (P)  34         6.7%
Ser (S)  32         6.3%
Thr (T)  33         6.5%
Trp (W)   7         1.4%
Tyr (Y)  18         3.5%
Val (V)  35         6.9%
```



Phe

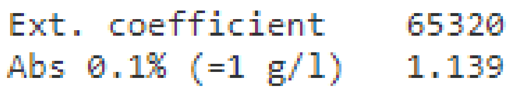


Trp

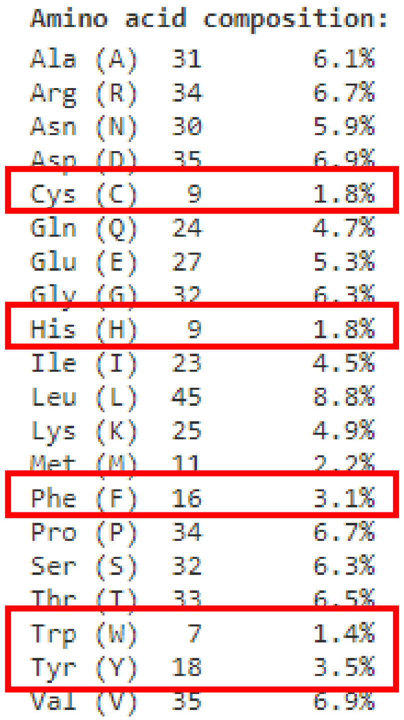


Tyr

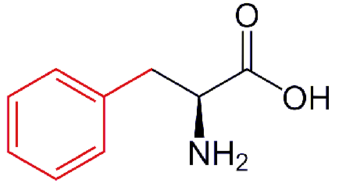
What determines the E value?



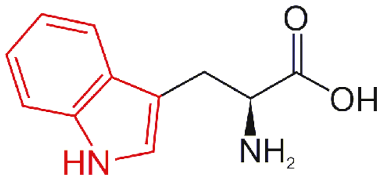
```
Ext. coefficient     65320
Abs 0.1% (=1 g/l)    1.139
```



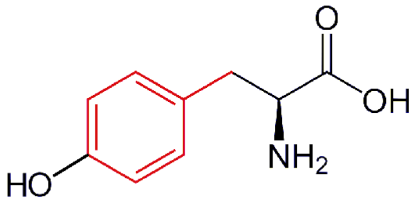
```
Amino acid composition:
Ala (A)  31          6.1%
Arg (R)  34          6.7%
Asn (N)  30          5.9%
Asp (D)  35          6.9%
Cys (C)   9          1.8%
Gln (Q)  24          4.7%
Glu (E)  27          5.3%
Gly (G)  32          6.3%
His (H)   9          1.8%
Ile (I)  23          4.5%
Leu (L)  45          8.8%
Lys (K)  25          4.9%
Met (M)  11          2.2%
Phe (F)  16          3.1%
Pro (P)  34          6.7%
Ser (S)  32          6.3%
Thr (T)  33          6.5%
Trp (W)   7          1.4%
Tyr (Y)  18          3.5%
Val (V)  35          6.9%
```



Phe



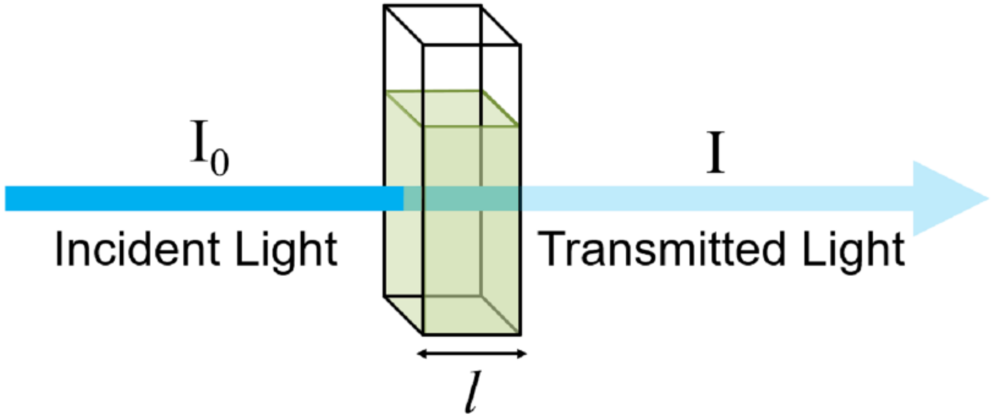
Trp



Tyr

Calculating protein concentration

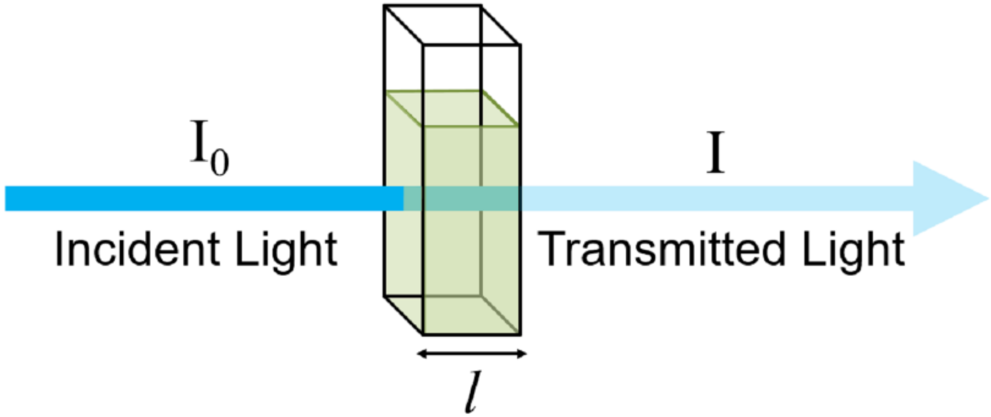
- Spectrophotometry, measure absorbance at 280 nm



$$A = log_{10}\left(\frac{I_0}{I}\right)$$

Calculating protein concentration

- Beer-Lambert Law

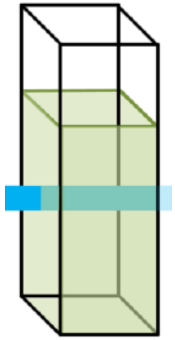


$$A = log_{10}\left(\frac{I_0}{I}\right) = \varepsilon c l$$

ε = extinction co-efficient
c = concentration
l = path length

Calculating protein concnetration

- Beer-Lambert Law and molecular extinction coefficients



$$A = \varepsilon c l$$

$$c = \frac{A}{\varepsilon l}$$

Measure the absorbance

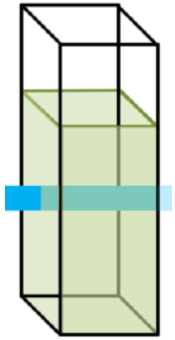
Use the theoretical extinction co-efficient

Calculate concentration

© Edinburgh Instruments

Calculating protein concentration

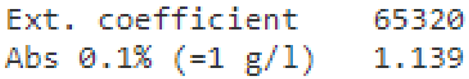
- Beer-Lambert Law and molecular extinction coefficients



$$A = 0.5$$

$$c = \frac{A}{\varepsilon l}$$

Calculate concentration in M or mg/mL

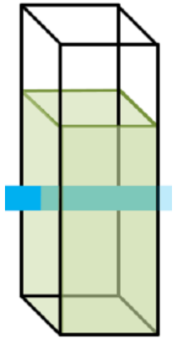


```
Ext. coefficient    65320
Abs 0.1% (=1 g/l)   1.139
```

© Edinburgh Instruments

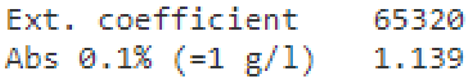
Calculating protein concentration

- Beer-Lambert Law and molecular extinction coefficients



$$A = 0.5$$

$$c = \frac{A}{\varepsilon l}$$



```
Ext. coefficient     65320
Abs 0.1% (=1 g/l)    1.139
```

Calculate concentration in M or mg/mL

Molar conc, 65320 (M)

$= 0.5/65320 = 7.65 \times 10^{-6}$

$= 7.65\ \mu M$

In mg/mL, 1.139

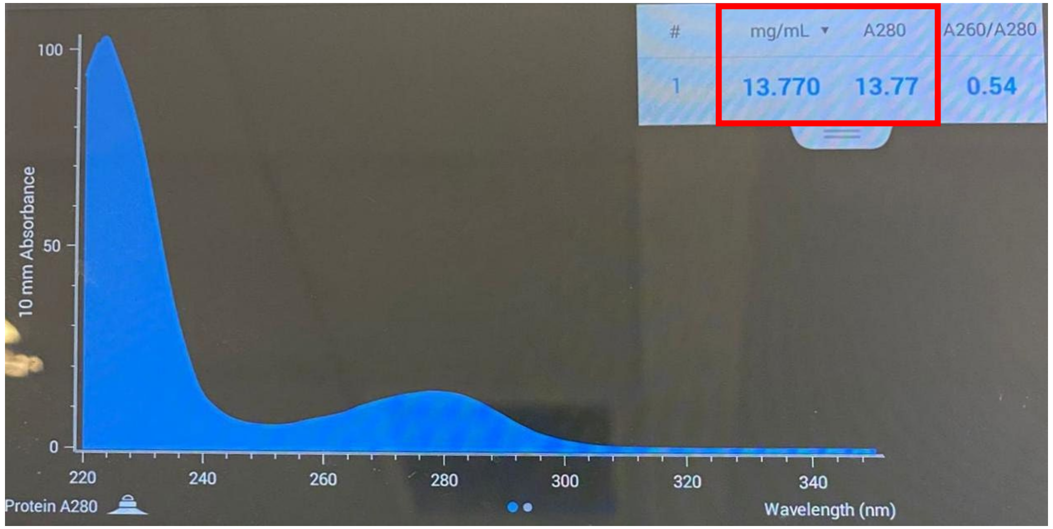
$= 0.5/1.139 = 0.44$ mg/mL

© Edinburgh Instruments

This correction is super important

- Always correct your absorbance measurements using your extinction coefficient

- Some instruments will quote concentration from absorbance but if you haven’t entered an extinction coefficient it will be wrong!



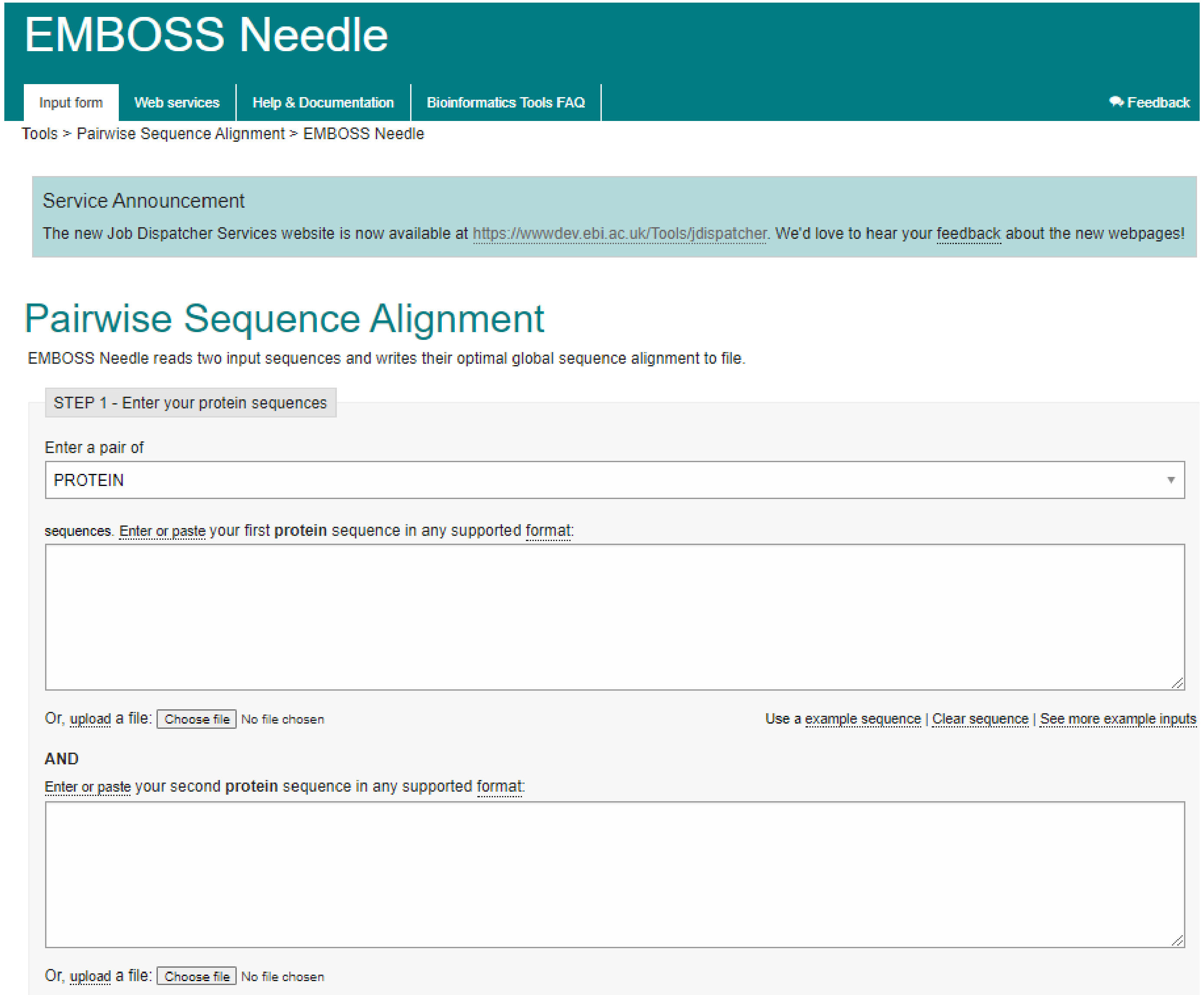
#	mg/mL ▾	A280	A260/A280
1	13.770	13.77	0.54

© Nanodrop, Thermo scientific

Sequence alignments

- Pairwise sequence alignments
  - mouse vs human
  - Proteins from different viruses

https://www.ebi.ac.uk/Tools/psa/emboss_needle/



EMBOSS Needle

Tools > Pairwise Sequence Alignment > EMBOSS Needle

Service Announcement

The new Job Dispatcher Services website is now available at https://wwwdev.ebi.ac.uk/Tools/jdispatcher. We'd love to hear your feedback about the new webpages!

Pairwise Sequence Alignment

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

STEP 1 - Enter your protein sequences

Enter a pair of PROTEIN sequences. Enter or paste your first **protein** sequence in any supported format:

Or, upload a file:

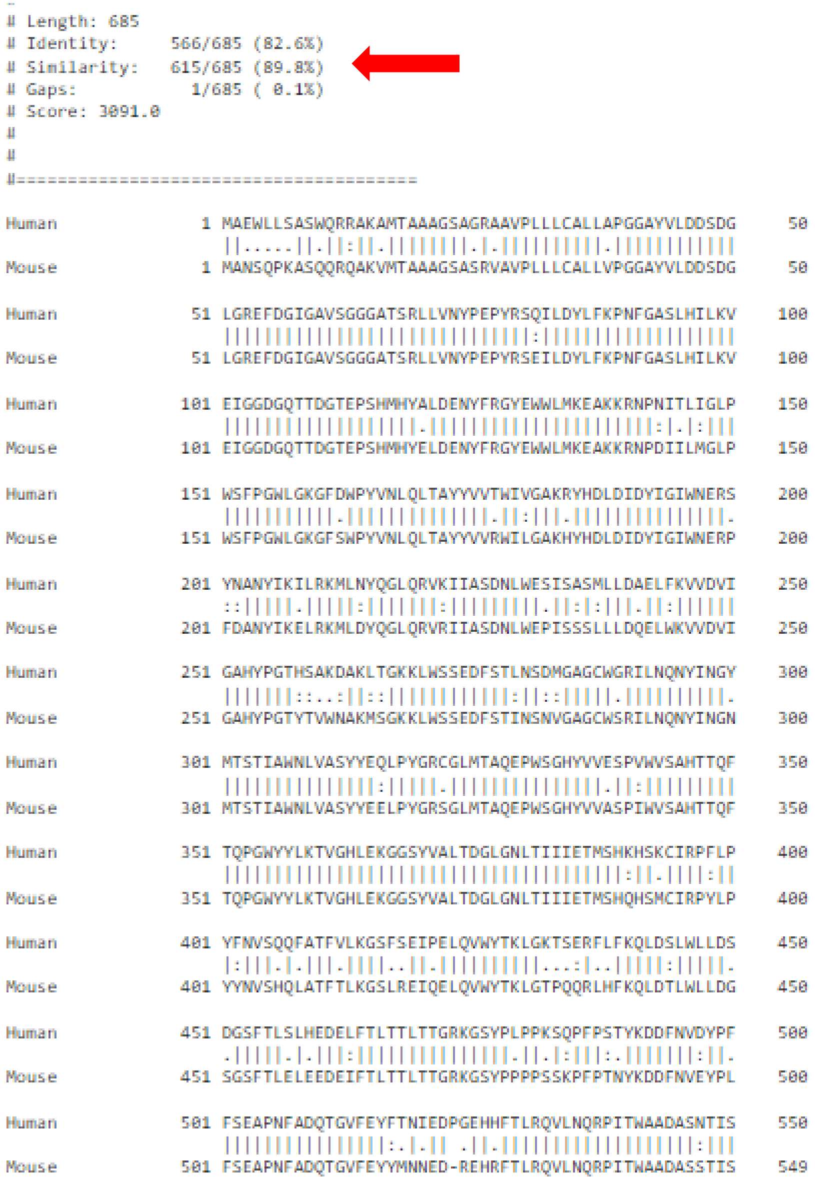
AND

Enter or paste your second **protein** sequence in any supported format:

Or, upload a file:

Pairwise alignment

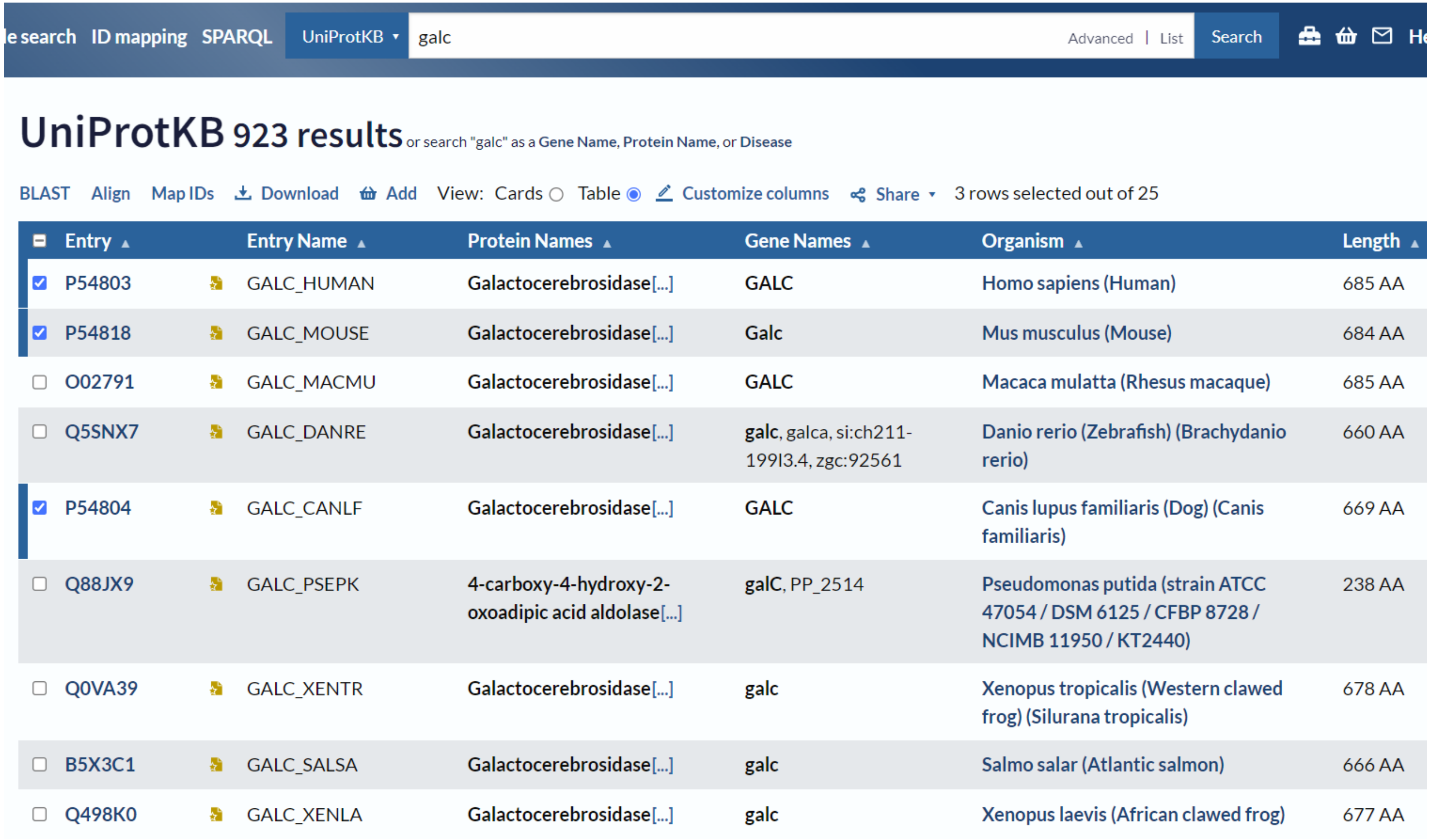
- Identity calculation
- Use similarity/homology carefully, calculated differently by programs
- Useful for comparing animal models of human diseases



```
# Length: 685
# Identity:     566/685 (82.6%)
# Similarity:   615/685 (89.8%)
# Gaps:           1/685 ( 0.1%)
# Score: 3091.0
#
#
#=======================================

Human              1 MAEWLLSASWQRRAKAMTAAAGSAGRAAVPLLLCALLAPGGAYVLDDSDG     50
                     ||.....||.||:||.||||||||.|.||||||||||.||||||||||||
Mouse              1 MANSQPKASQQRQAKVMTAAAGSASRVAVPLLLCALLVPGGAYVLDDSDG     50

Human             51 LGREFDGIGAVSGGGATSRLLVNYPEPYRSQILDYLFKPNFGASLHILKV    100
                     ||||||||||||||||||||||||||||||:|||||||||||||||||||
Mouse             51 LGREFDGIGAVSGGGATSRLLVNYPEPYRSEILDYLFKPNFGASLHILKV    100

Human            101 EIGGDGQTTDGTEPSHMHYALDENYFRGYEWWLMKEAKKRNPNITLIGLP    150
                     |||||||||||||||||||.||||||||||||||||||||||:|.|:|||
Mouse            101 EIGGDGQTTDGTEPSHMHYELDENYFRGYEWWLMKEAKKRNPDIILMGLP    150

Human            151 WSFPGWLGKGFDWPYVNLQLTAYYVVTWIVGAKRYHDLDIDYIGIWNERS    200
                     |||||||||||.||||||||||||||.||:|||.|||||||||||||||.
Mouse            151 WSFPGWLGKGFSWPYVNLQLTAYYVVRWILGAKHYHDLDIDYIGIWNERP    200

Human            201 YNANYIKILRKMLNYQGLQRVKIIASDNLWESISASMLLDAELFKVVDVI    250
                     ::|||||.|||||:|||||||:|||||||||.||:|:|||.||:||||||
Mouse            201 FDANYIKELRKMLDYQGLQRVRIIASDNLWEPISSSLLLDQELWKVVDVI    250

Human            251 GAHYPGTHSAKDAKLTGKKLWSSEDFSTLNSDMGAGCWGRILNQNYINGY    300
                     |||||||::..:||::||||||||||||:||::|||||.||||||||||.
Mouse            251 GAHYPGTYTVWNAKMSGKKLWSSEDFSTINSNVGAGCWSRILNQNYINGN    300

Human            301 MTSTIAWNLVASYYEQLPYGRCGLMTAQEPWSGHYVVESPVWVSAHTTQF    350
                     |||||||||||||||:|||||.|||||||||||||||.||:|||||||||
Mouse            301 MTSTIAWNLVASYYEELPYGRSGLMTAQEPWSGHYVVASPIWVSAHTTQF    350

Human            351 TQPGWYYLKTVGHLEKGGSYVALTDGLGNLTIIIETMSHKHSKCIRPFLP    400
                     |||||||||||||||||||||||||||||||||||||||:||.||||:||
Mouse            351 TQPGWYYLKTVGHLEKGGSYVALTDGLGNLTIIIETMSHQHSMCIRPYLP    400

Human            401 YFNVSQQFATFVLKGSFSEIPELQVWYTKLGKTSERFLFKQLDSLWLLDS    450
                     |:|||.|.|||.||||..||.||||||||||...:|..|||||:|||||.
Mouse            401 YYNVSHQLATFTLKGSLREIQELQVWYTKLGTPQQRLHFKQLDTLWLLDG    450

Human            451 DGSFTLSLHEDELFTLTTLTTGRKGSYPLPPKSQPFPSTYKDDFNVDYPF    500
                     .|||||.|.|||:|||||||||||||||.||.|:|||:.|||||||:||.
Mouse            451 SGSFTLELEEDEIFTLTTLTTGRKGSYPPPPSSKPFPTNYKDDFNVEYPL    500

Human            501 FSEAPNFADQTGVFEYFTNIEDPGEHHFTLRQVLNQRPITWAADASNTIS    550
                     ||||||||||||||||:.|.|| .||.|||||||||||||||||||:|||
Mouse            501 FSEAPNFADQTGVFEYYMNNED-REHRFTLRQVLNQRPITWAADASSTIS    549
```

You can also use UniProt to do alignments

- Search for your gene/protein of interest
- Select the ones you want to compare
- Click Align

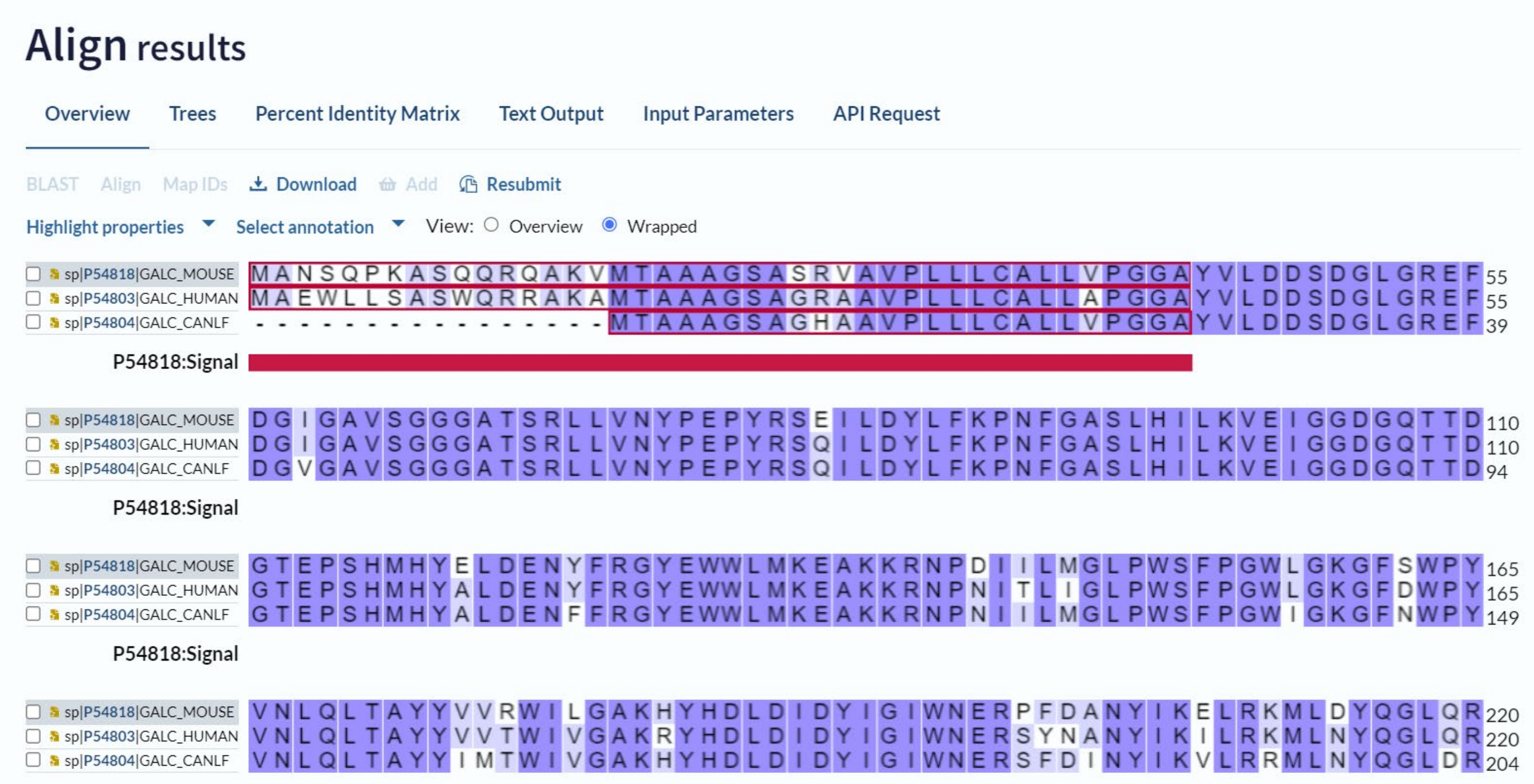


UniProtKB 923 results or search "galc" as a Gene Name, Protein Name, or Disease

View: Cards ○ Table ● Customize columns Share ▾ 3 rows selected out of 25

☐	Entry ▲	Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
☑	P54803	GALC_HUMAN	Galactocerebrosidase[...]	GALC	Homo sapiens (Human)	685 AA
☑	P54818	GALC_MOUSE	Galactocerebrosidase[...]	Galc	Mus musculus (Mouse)	684 AA
☐	O02791	GALC_MACMU	Galactocerebrosidase[...]	GALC	Macaca mulatta (Rhesus macaque)	685 AA
☐	Q5SNX7	GALC_DANRE	Galactocerebrosidase[...]	galc, galca, si:ch211-199I3.4, zgc:92561	Danio rerio (Zebrafish) (Brachydanio rerio)	660 AA
☑	P54804	GALC_CANLF	Galactocerebrosidase[...]	GALC	Canis lupus familiaris (Dog) (Canis familiaris)	669 AA
☐	Q88JX9	GALC_PSEPK	4-carboxy-4-hydroxy-2-oxoadipic acid aldolase[...]	galC, PP_2514	Pseudomonas putida (strain ATCC 47054 / DSM 6125 / CFBP 8728 / NCIMB 11950 / KT2440)	238 AA
☐	Q0VA39	GALC_XENTR	Galactocerebrosidase[...]	galc	Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)	678 AA
☐	B5X3C1	GALC_SALSA	Galactocerebrosidase[...]	galc	Salmo salar (Atlantic salmon)	666 AA
☐	Q498K0	GALC_XENLA	Galactocerebrosidase[...]	galc	Xenopus laevis (African clawed frog)	677 AA

UniProt Alignment



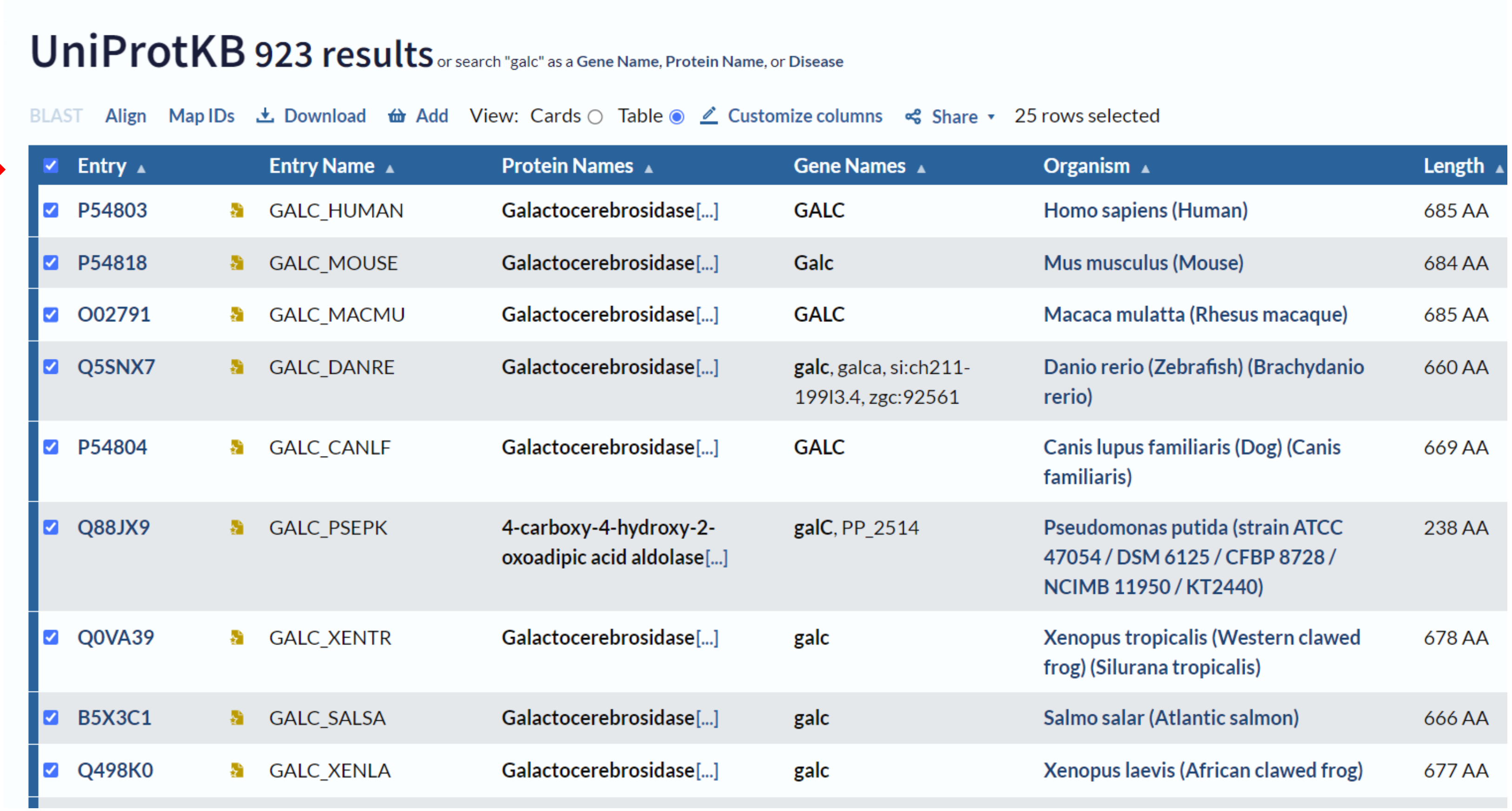
Align results

View: ○ Overview ● Wrapped

```
□ sp|P54818|GALC_MOUSE  MANSQPKASQQRQAKVMTAAAGSASRVAVPLLLCALLVPGGAYVLDDSDGLGREF 55
□ sp|P54803|GALC_HUMAN  MAEWLLSASWQRRAKAMTAAAGSAGRAAVPLLLCALLAPGGAYVLDDSDGLGREF 55
□ sp|P54804|GALC_CANLF  ----------------MTAAAGSAGHAAVPLLLCALLVPGGAYVLDDSDGLGREF 39
          P54818:Signal

□ sp|P54818|GALC_MOUSE  DGIGAVSGGGATSRLLVNYPEPYRSEILDYLFKPNFGASLHILKVEIGGDGQTTD 110
□ sp|P54803|GALC_HUMAN  DGIGAVSGGGATSRLLVNYPEPYRSQILDYLFKPNFGASLHILKVEIGGDGQTTD 110
□ sp|P54804|GALC_CANLF  DGVGAVSGGGATSRLLVNYPEPYRSQILDYLFKPNFGASLHILKVEIGGDGQTTD 94
          P54818:Signal

□ sp|P54818|GALC_MOUSE  GTEPSHMHYELDENYFRGYEWWLMKEAKKRNPDIILMGLPWSFPGWLGKGFSWPY 165
□ sp|P54803|GALC_HUMAN  GTEPSHMHYALDENYFRGYEWWLMKEAKKRNPNITLIGLPWSFPGWLGKGFDWPY 165
□ sp|P54804|GALC_CANLF  GTEPSHMHYALDENFFRGYEWWLMKEAKKRNPNIILMGLPWSFPGWIGKGFNWPY 149
          P54818:Signal

□ sp|P54818|GALC_MOUSE  VNLQLTAYYVVRWILGAKHYHDLDIDYIGIWNERPFDANYIKELRKMLDYQGLQR 220
□ sp|P54803|GALC_HUMAN  VNLQLTAYYVVTWIVGAKRYHDLDIDYIGIWNERSYNANYIKILRKMLNYQGLQR 220
□ sp|P54804|GALC_CANLF  VNLQLTAYYIMTWIVGAKHYHDLDIDYIGIWNERSFDINYIKVLRRMLNYQGLDR 204
```

You can also make phylogenetic trees

- Select all genes
- Do Alignment

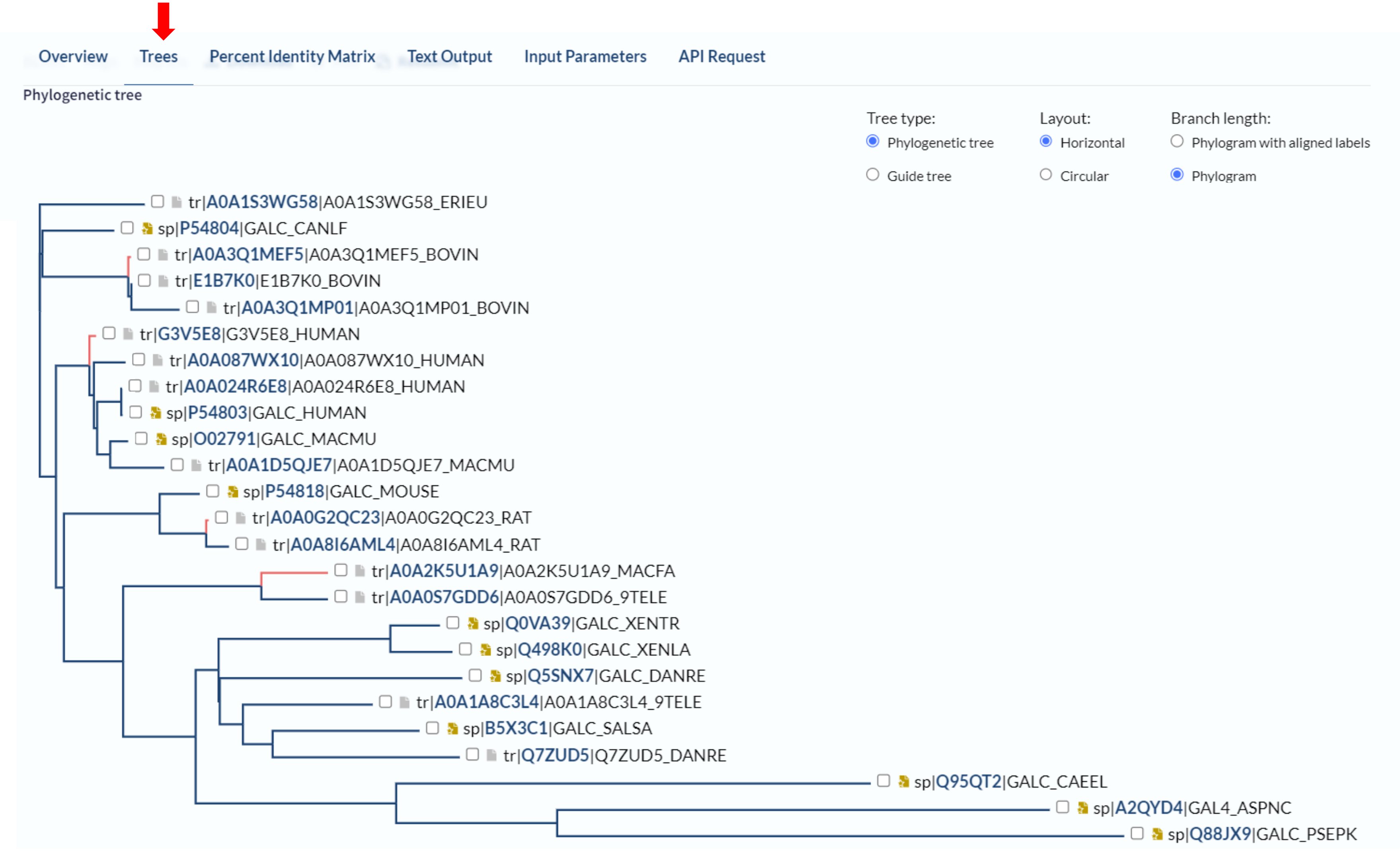


UniProtKB 923 results or search "galc" as a Gene Name, Protein Name, or Disease

☑ Entry ▲	Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
☑ P54803	GALC_HUMAN	Galactocerebrosidase[...]	GALC	Homo sapiens (Human)	685 AA
☑ P54818	GALC_MOUSE	Galactocerebrosidase[...]	Galc	Mus musculus (Mouse)	684 AA
☑ O02791	GALC_MACMU	Galactocerebrosidase[...]	GALC	Macaca mulatta (Rhesus macaque)	685 AA
☑ Q5SNX7	GALC_DANRE	Galactocerebrosidase[...]	galc, galca, si:ch211-199l3.4, zgc:92561	Danio rerio (Zebrafish) (Brachydanio rerio)	660 AA
☑ P54804	GALC_CANLF	Galactocerebrosidase[...]	GALC	Canis lupus familiaris (Dog) (Canis familiaris)	669 AA
☑ Q88JX9	GALC_PSEPK	4-carboxy-4-hydroxy-2-oxoadipic acid aldolase[...]	galC, PP_2514	Pseudomonas putida (strain ATCC 47054 / DSM 6125 / CFBP 8728 / NCIMB 11950 / KT2440)	238 AA
☑ Q0VA39	GALC_XENTR	Galactocerebrosidase[...]	galc	Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)	678 AA
☑ B5X3C1	GALC_SALSA	Galactocerebrosidase[...]	galc	Salmo salar (Atlantic salmon)	666 AA
☑ Q498K0	GALC_XENLA	Galactocerebrosidase[...]	galc	Xenopus laevis (African clawed frog)	677 AA

You can also make phylogenetic trees

- Select all genes
- Do Alignment
- Select Trees



Tree type: ● Phylogenetic tree ○ Guide tree

Layout: ● Horizontal ○ Circular

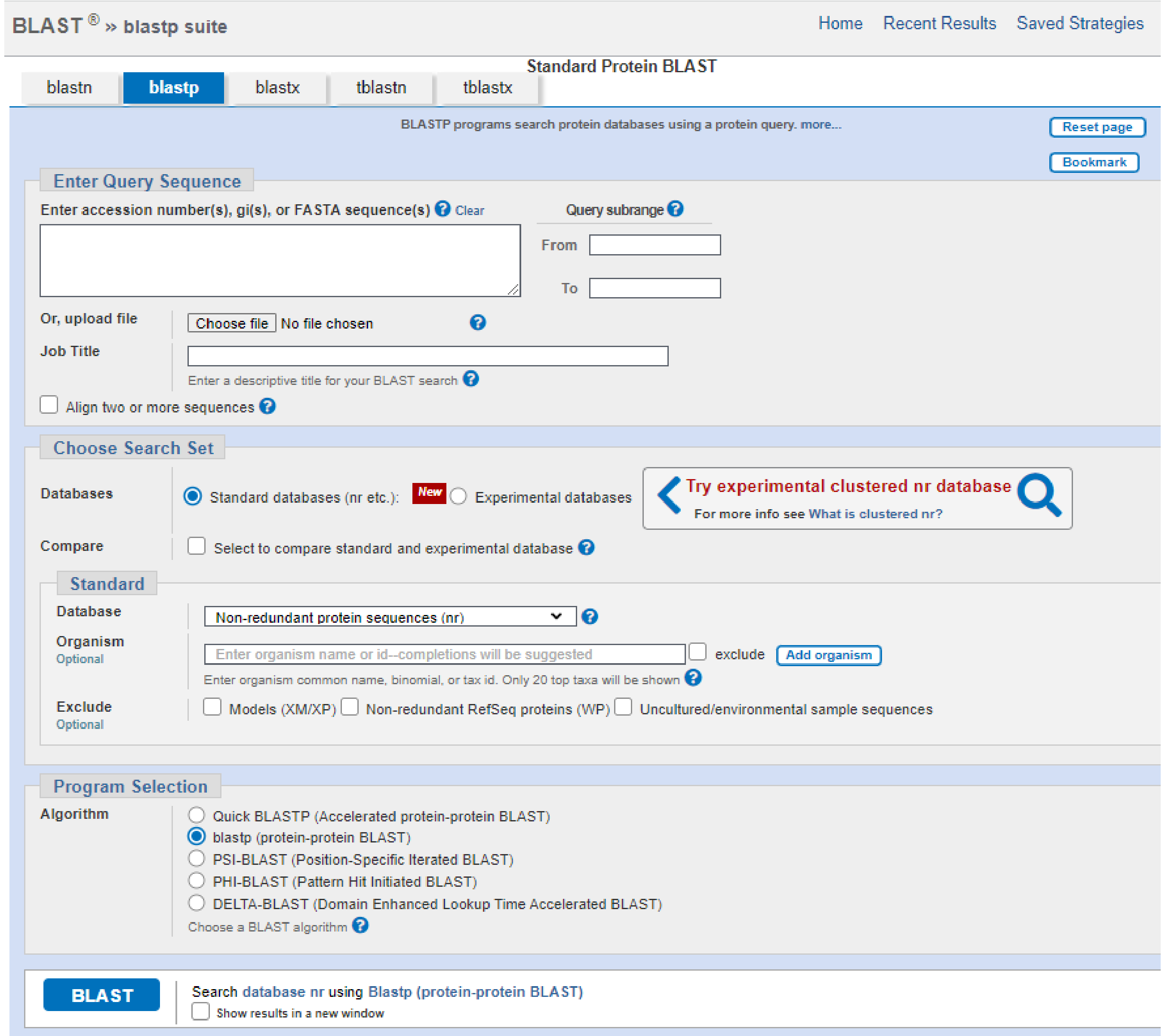
Branch length: ○ Phylogram with aligned labels ● Phylogram

□ tr|A0A1S3WG58|A0A1S3WG58_ERIEU
□ sp|P54804|GALC_CANLF
□ tr|A0A3Q1MEF5|A0A3Q1MEF5_BOVIN
□ tr|E1B7K0|E1B7K0_BOVIN
□ tr|A0A3Q1MP01|A0A3Q1MP01_BOVIN
□ tr|G3V5E8|G3V5E8_HUMAN
□ tr|A0A087WX10|A0A087WX10_HUMAN
□ tr|A0A024R6E8|A0A024R6E8_HUMAN
□ sp|P54803|GALC_HUMAN
□ sp|O02791|GALC_MACMU
□ tr|A0A1D5QJE7|A0A1D5QJE7_MACMU
□ sp|P54818|GALC_MOUSE
□ tr|A0A0G2QC23|A0A0G2QC23_RAT
□ tr|A0A8I6AML4|A0A8I6AML4_RAT
□ tr|A0A2K5U1A9|A0A2K5U1A9_MACFA
□ tr|A0A0S7GDD6|A0A0S7GDD6_9TELE
□ sp|Q0VA39|GALC_XENTR
□ sp|Q498K0|GALC_XENLA
□ sp|Q5SNX7|GALC_DANRE
□ tr|A0A1A8C3L4|A0A1A8C3L4_9TELE
□ sp|B5X3C1|GALC_SALSA
□ tr|Q7ZUD5|Q7ZUD5_DANRE
□ sp|Q95QT2|GALC_CAEEL
□ sp|A2QYD4|GAL4_ASPNC
□ sp|Q88JX9|GALC_PSEPK

Sequence homology

- Identify conservation across species

- Highlight important functional regions of high conservation



BLAST ® » blastp suite

Home Recent Results Saved Strategies

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

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

Query subrange ?

From

To

Or, upload file: Choose file No file chosen ?

Job Title

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database ?

Standard

Database: Non-redundant protein sequences (nr) ?

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) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

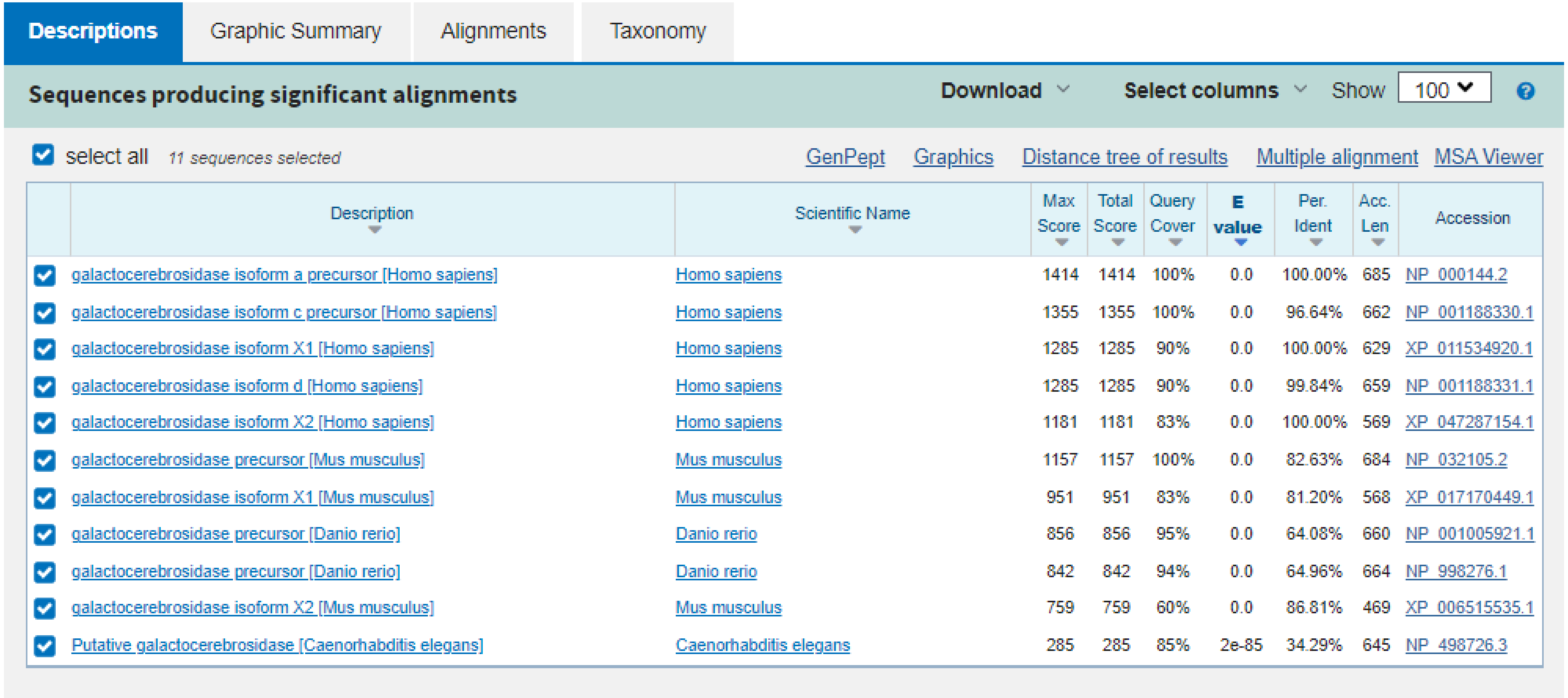
Choose a BLAST algorithm ?

BLAST

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

Sequence homology

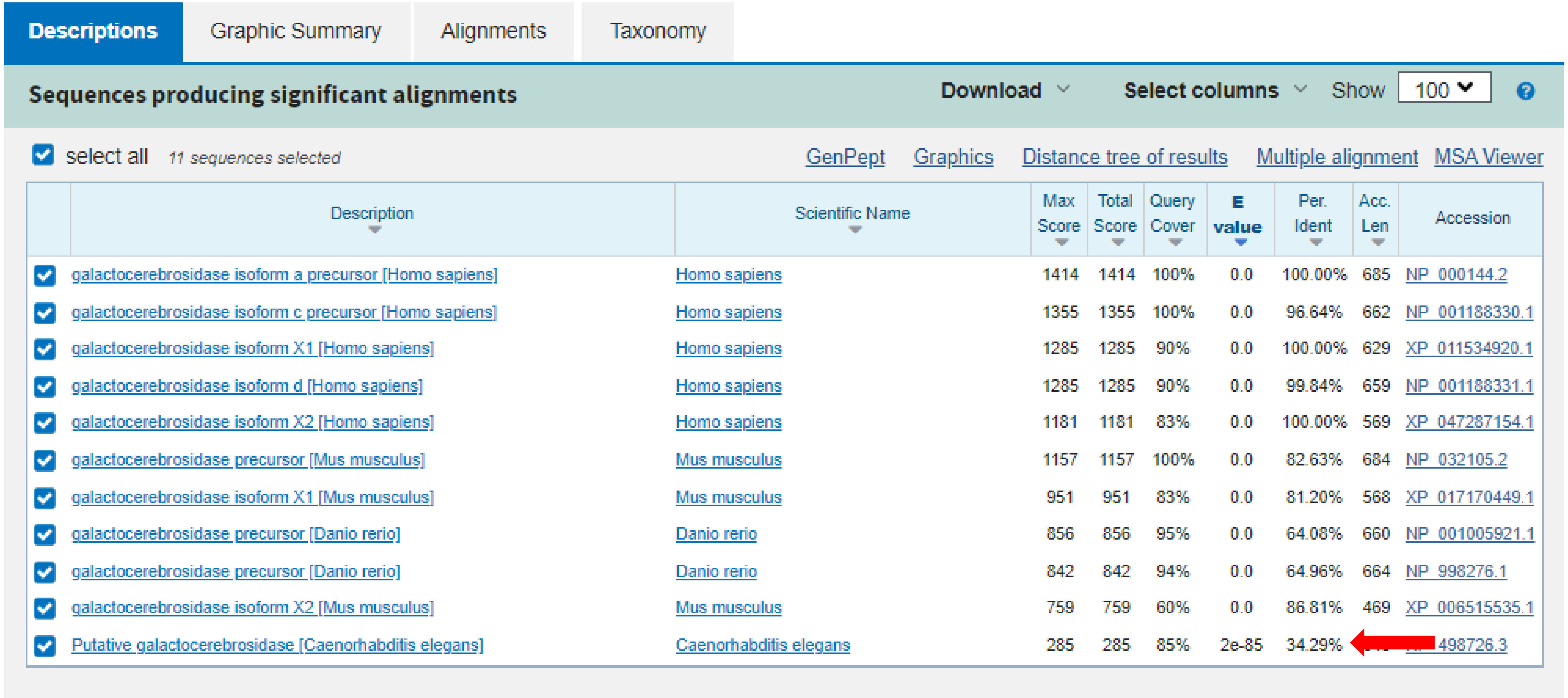


Sequences producing significant alignments

☑ select all *11 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	galactocerebrosidase isoform a precursor [Homo sapiens]	Homo sapiens	1414	1414	100%	0.0	100.00%	685	NP_000144.2
☑	galactocerebrosidase isoform c precursor [Homo sapiens]	Homo sapiens	1355	1355	100%	0.0	96.64%	662	NP_001188330.1
☑	galactocerebrosidase isoform X1 [Homo sapiens]	Homo sapiens	1285	1285	90%	0.0	100.00%	629	XP_011534920.1
☑	galactocerebrosidase isoform d [Homo sapiens]	Homo sapiens	1285	1285	90%	0.0	99.84%	659	NP_001188331.1
☑	galactocerebrosidase isoform X2 [Homo sapiens]	Homo sapiens	1181	1181	83%	0.0	100.00%	569	XP_047287154.1
☑	galactocerebrosidase precursor [Mus musculus]	Mus musculus	1157	1157	100%	0.0	82.63%	684	NP_032105.2
☑	galactocerebrosidase isoform X1 [Mus musculus]	Mus musculus	951	951	83%	0.0	81.20%	568	XP_017170449.1
☑	galactocerebrosidase precursor [Danio rerio]	Danio rerio	856	856	95%	0.0	64.08%	660	NP_001005921.1
☑	galactocerebrosidase precursor [Danio rerio]	Danio rerio	842	842	94%	0.0	64.96%	664	NP_998276.1
☑	galactocerebrosidase isoform X2 [Mus musculus]	Mus musculus	759	759	60%	0.0	86.81%	469	XP_006515535.1
☑	Putative galactocerebrosidase [Caenorhabditis elegans]	Caenorhabditis elegans	285	285	85%	2e-85	34.29%	645	NP_498726.3

Sequence homology

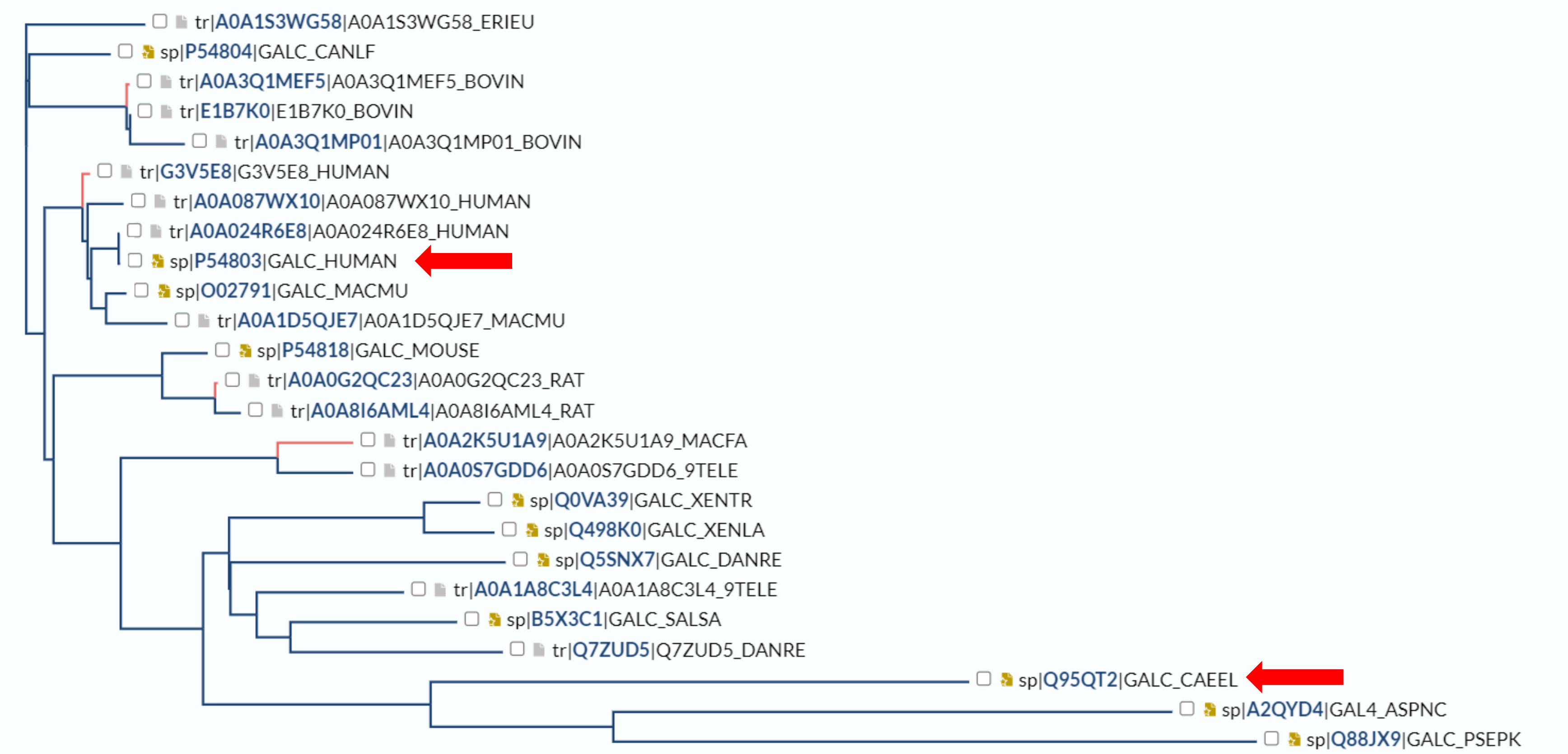


Sequences producing significant alignments

☑ select all *11 sequences selected*

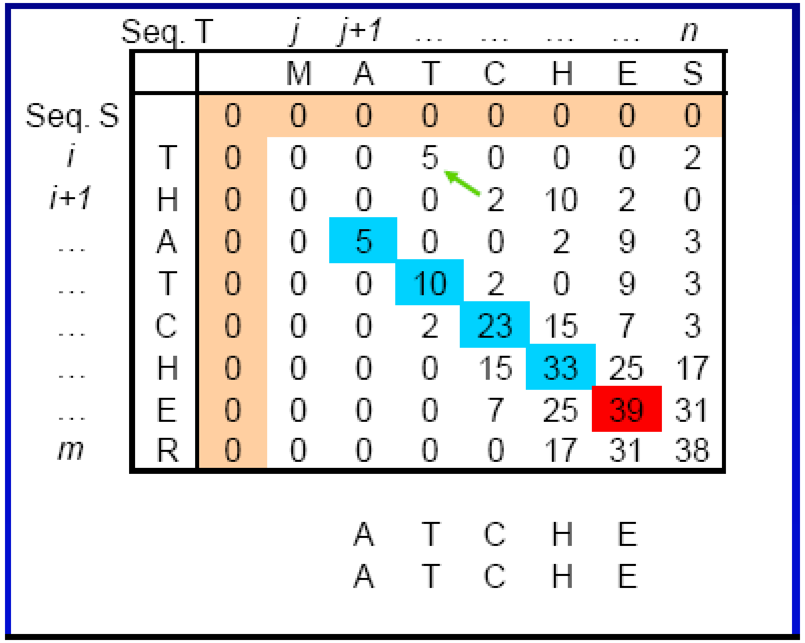
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	galactocerebrosidase isoform a precursor [Homo sapiens]	Homo sapiens	1414	1414	100%	0.0	100.00%	685	NP_000144.2
☑	galactocerebrosidase isoform c precursor [Homo sapiens]	Homo sapiens	1355	1355	100%	0.0	96.64%	662	NP_001188330.1
☑	galactocerebrosidase isoform X1 [Homo sapiens]	Homo sapiens	1285	1285	90%	0.0	100.00%	629	XP_011534920.1
☑	galactocerebrosidase isoform d [Homo sapiens]	Homo sapiens	1285	1285	90%	0.0	99.84%	659	NP_001188331.1
☑	galactocerebrosidase isoform X2 [Homo sapiens]	Homo sapiens	1181	1181	83%	0.0	100.00%	569	XP_047287154.1
☑	galactocerebrosidase precursor [Mus musculus]	Mus musculus	1157	1157	100%	0.0	82.63%	684	NP_032105.2
☑	galactocerebrosidase isoform X1 [Mus musculus]	Mus musculus	951	951	83%	0.0	81.20%	568	XP_017170449.1
☑	galactocerebrosidase precursor [Danio rerio]	Danio rerio	856	856	95%	0.0	64.08%	660	NP_001005921.1
☑	galactocerebrosidase precursor [Danio rerio]	Danio rerio	842	842	94%	0.0	64.96%	664	NP_998276.1
☑	galactocerebrosidase isoform X2 [Mus musculus]	Mus musculus	759	759	60%	0.0	86.81%	469	XP_006515535.1
☑	Putative galactocerebrosidase [Caenorhabditis elegans]	Caenorhabditis elegans	285	285	85%	2e-85	34.29%	[illegible]	NP_498726.3

Comparing Outputs – trees and alignments


□ tr|A0A1S3WG58|A0A1S3WG58_ERIEU
□ sp|P54804|GALC_CANLF
□ tr|A0A3Q1MEF5|A0A3Q1MEF5_BOVIN
□ tr|E1B7K0|E1B7K0_BOVIN
□ tr|A0A3Q1MP01|A0A3Q1MP01_BOVIN
□ tr|G3V5E8|G3V5E8_HUMAN
□ tr|A0A087WX10|A0A087WX10_HUMAN
□ tr|A0A024R6E8|A0A024R6E8_HUMAN
□ sp|P54803|GALC_HUMAN
□ sp|O02791|GALC_MACMU
□ tr|A0A1D5QJE7|A0A1D5QJE7_MACMU
□ sp|P54818|GALC_MOUSE
□ tr|A0A0G2QC23|A0A0G2QC23_RAT
□ tr|A0A8I6AML4|A0A8I6AML4_RAT
□ tr|A0A2K5U1A9|A0A2K5U1A9_MACFA
□ tr|A0A0S7GDD6|A0A0S7GDD6_9TELE
□ sp|Q0VA39|GALC_XENTR
□ sp|Q498K0|GALC_XENLA
□ sp|Q5SNX7|GALC_DANRE
□ tr|A0A1A8C3L4|A0A1A8C3L4_9TELE
□ sp|B5X3C1|GALC_SALSA
□ tr|Q7ZUD5|Q7ZUD5_DANRE
□ sp|Q95QT2|GALC_CAEEL
□ sp|A2QYD4|GAL4_ASPNC
□ sp|Q88JX9|GALC_PSEPK

Including conservation improves search sensitivity

- Standard sequence alignment maximises the correspondence of residues across both proteins



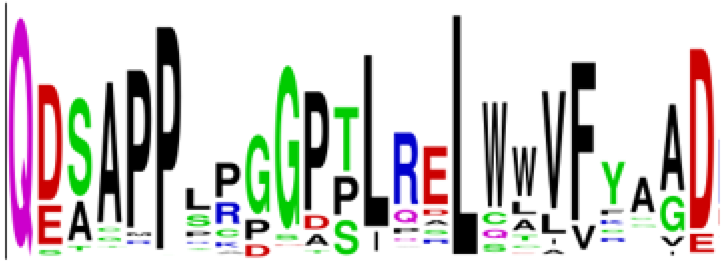
Seq. T			j	j+1	...	...	...	...	n
			M	A	T	C	H	E	S
Seq. S		0	0	0	0	0	0	0	0
i	T	0	0	0	5	0	0	0	2
i+1	H	0	0	0	0	2	10	2	0
...	A	0	0	5	0	0	2	9	3
...	T	0	0	0	10	2	0	9	3
...	C	0	0	0	2	23	15	7	3
...	H	0	0	0	0	15	33	25	17
...	E	0	0	0	0	7	25	39	31
m	R	0	0	0	0	0	17	31	38

```
A T C H E
A T C H E
```

Image: http://compbio.pbworks.com/w/page/16252912/Pairwise%20Sequence%20Alignment

Including conservation improves search sensitivity

- Standard sequence alignment maximises the correspondence of residues across both proteins
- But some residues will be *evolutionarily conserved* while others won’t
  - Conserved residues more likely to be important for function
- Weighting the alignment by conservation makes it more sensitive and accurate
  - Profiles and Hidden Markov Models (HMMs)

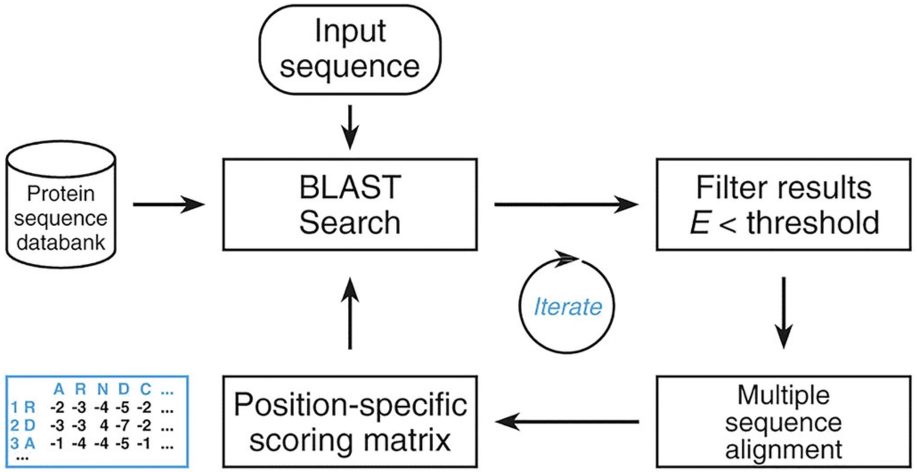


Logo Plot

Image: Stephen Graham (CC BY 4.0)

Profile/HMM searching for identifying distant homologs

- Use the *query sequence* to perform an initial search of the *sequence database*
- Take all hits from initial search and build a profile or HMM
- Use this profile/HMM to perform a subsequent search of the *same sequence database*
    - Will be more sensitive and accurate
- Profile searching: PSI-BLAST
- HMM searching: HMMER (HMMsearch)



Input sequence

Protein sequence databank

BLAST Search

Filter results E < threshold

Iterate

Multiple sequence alignment

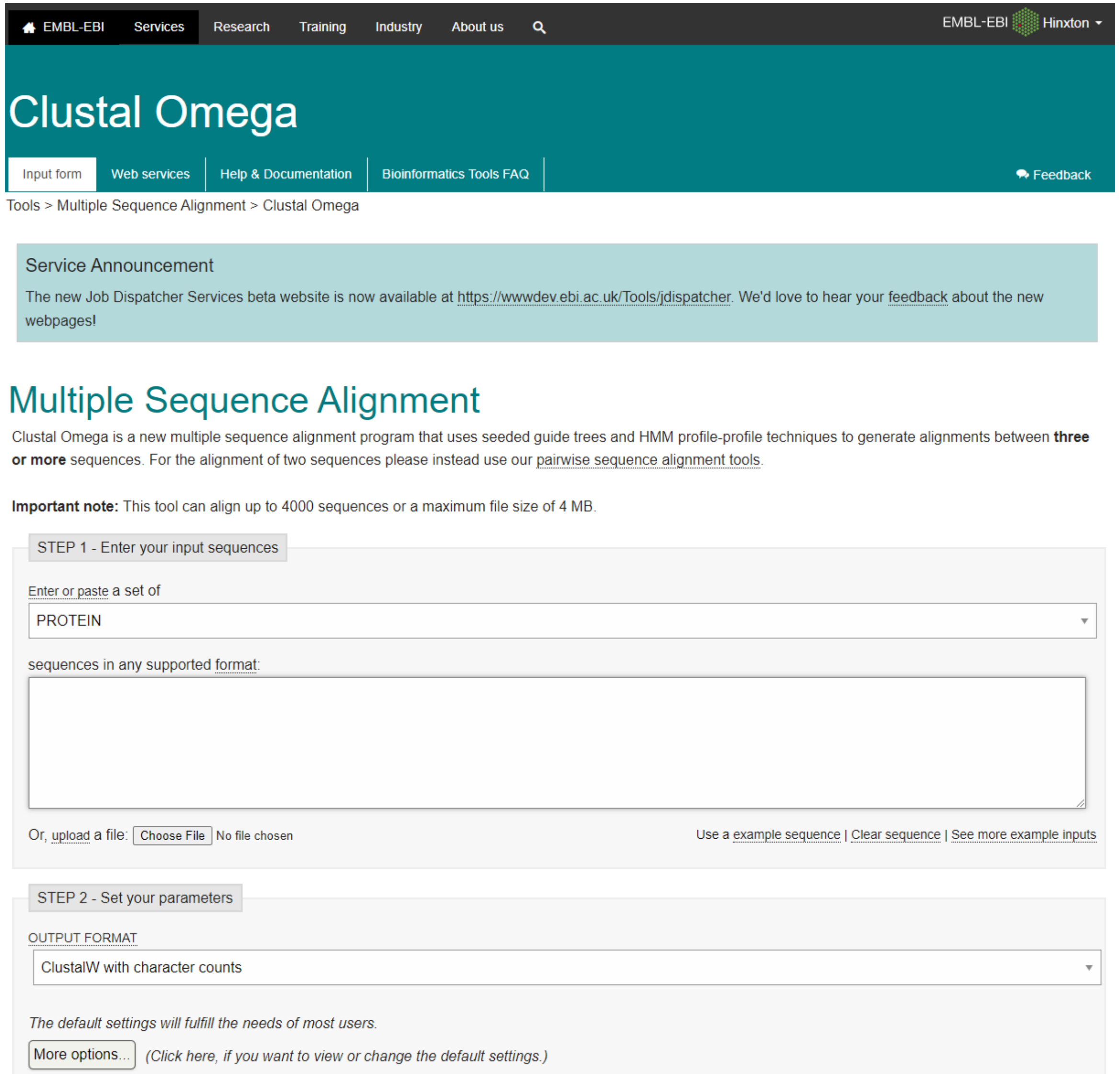
Position-specific scoring matrix

	A	R	N	D	C	...
1 R	-2	-3	-4	-5	-2	...
2 D	-3	-3	4	-7	-2	...
3 A	-1	-4	-4	-5	-1	...
...						

Image: Yousaf et al (2021) *International Review of Applied Sciences and Engineering* 13, 1

HMM:HMM alignments

- Can also use HMM/HMM comparisons to improve specific multiple sequence alignments
  - Upweight the alignment of regions that are evolutionarily conserved
- Clustal Omega



Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

Service Announcement

The new Job Dispatcher Services beta website is now available at https://wwwdev.ebi.ac.uk/Tools/jdispatcher. We'd love to hear your feedback about the new webpages!

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three 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

sequences in any supported format:

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

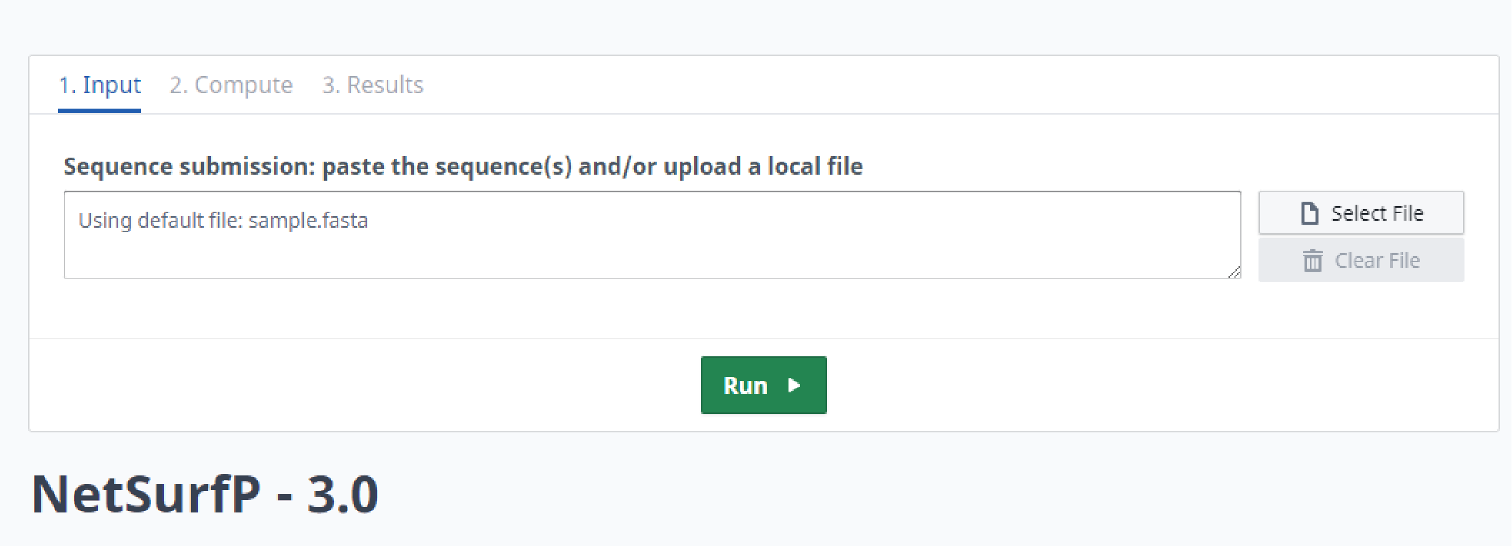
More options... *(Click here, if you want to view or change the default settings.)*

https://www.ebi.ac.uk/Tools/msa/clustalo/

What else can we predict from sequence?

- Secondary structure prediction using NetSurfP

https://dtu.biolib.com/NetSurfP-3/

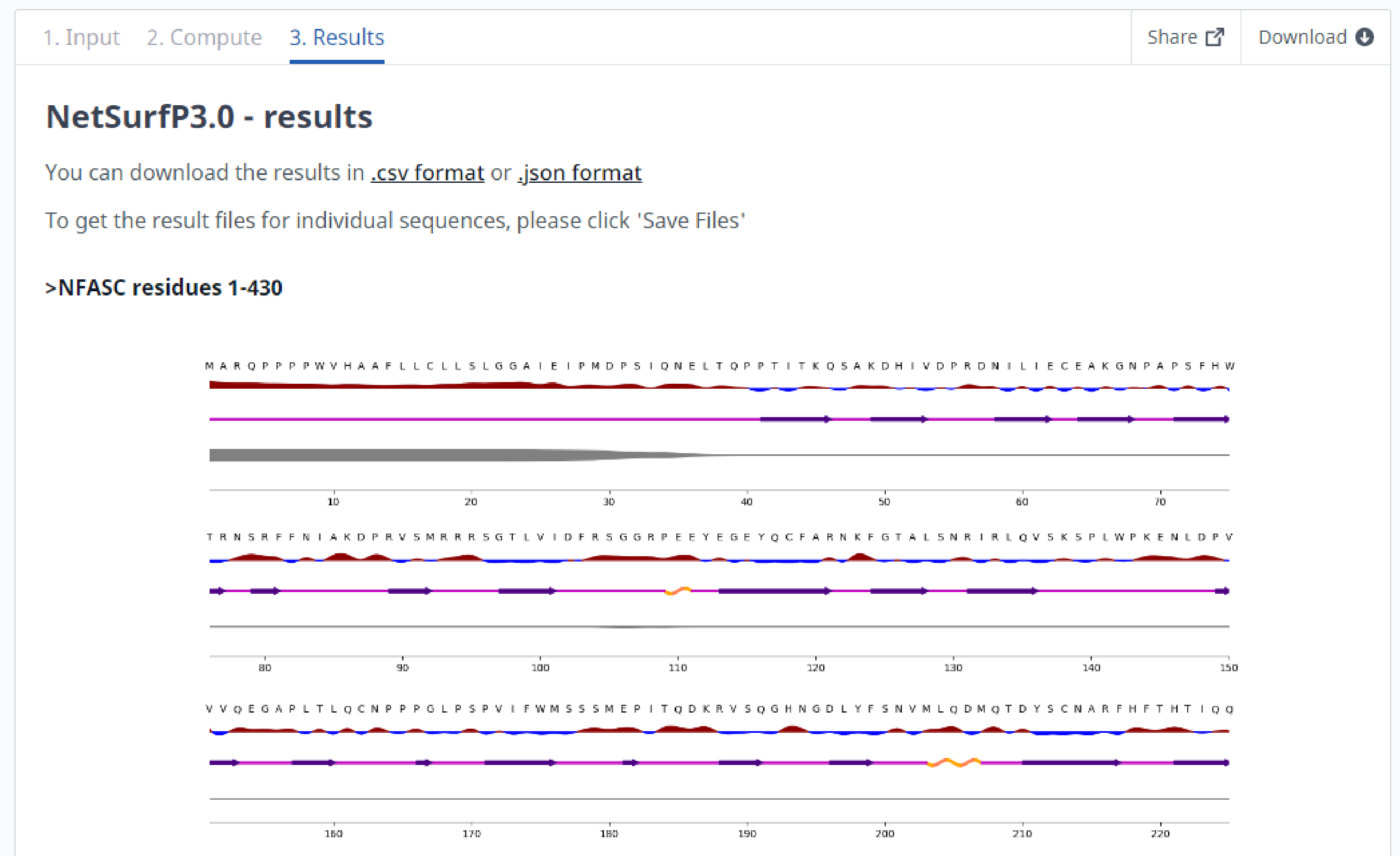


Sequence submission: paste the sequence(s) and/or upload a local file

Using default file: sample.fasta

NetSurfP - 3.0

Secondary structure and disorder



1. Input 2. Compute 3. Results

Share Download

NetSurfP3.0 - results

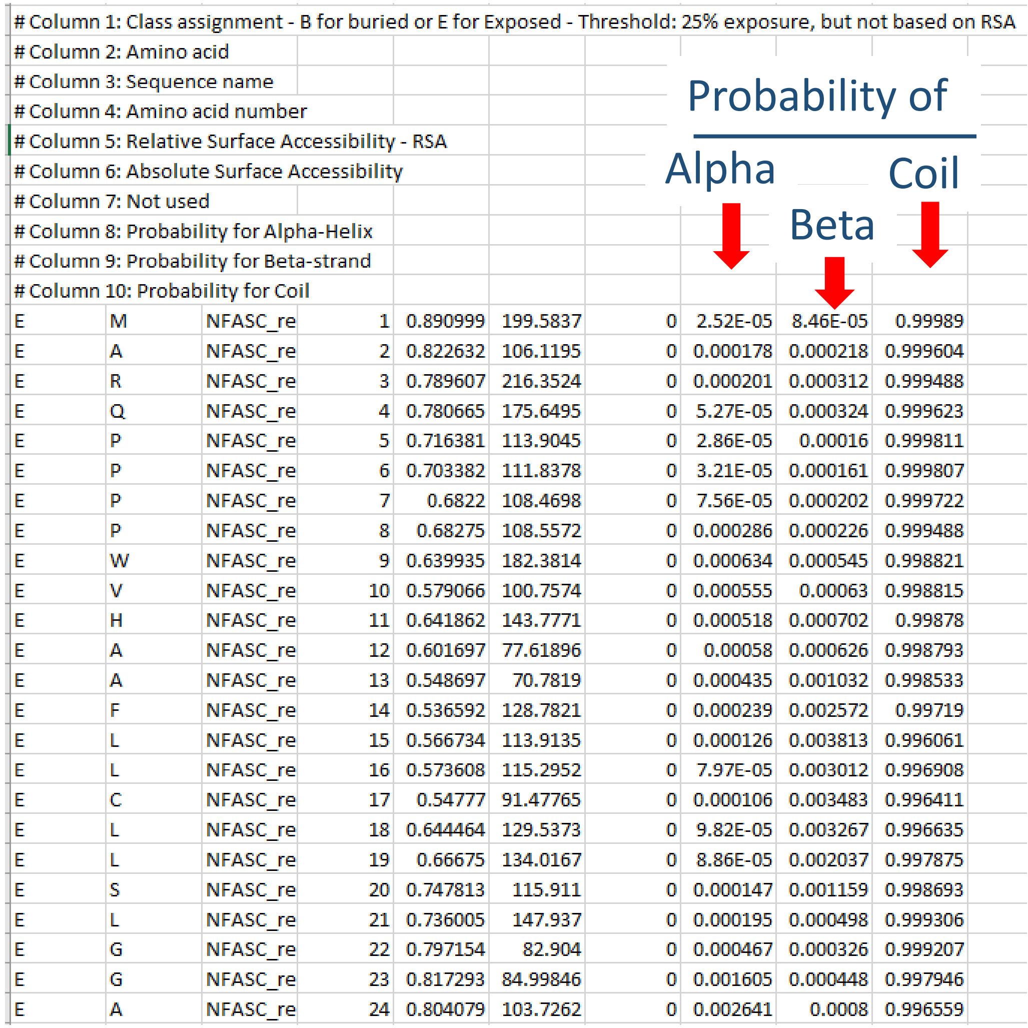
You can download the results in .csv format or .json format

To get the result files for individual sequences, please click 'Save Files'

>NFASC residues 1-430

M A R Q P P P P W V H A A F L L C L L S L G G A I E I P M D P S I Q N E L T Q P P T I T K Q S A K D H I V D P R D N I L I E C E A K G N P A P S F H W
10 20 30 40 50 60 70
T R N S R F F N I A K D P R V S M R R R S G T L V I D F R S G G R P E E Y E G E Y Q C F A R N K F G T A L S N R I R L Q V S K S P L W P K E N L D P V
80 90 100 110 120 130 140 150
V V Q E G A P L T L Q C N P P P G L P S P V I F W M S S S M E P I T Q D K R V S Q G H N G D L Y F S N V M L Q D M Q T D Y S C N A R F H F T H T I Q Q
160 170 180 190 200 210 220

Secondary structure and disorder



Column 1: Class assignment - B for buried or E for Exposed - Threshold: 25% exposure, but not based on RSA
Column 2: Amino acid
Column 3: Sequence name
Column 4: Amino acid number
Column 5: Relative Surface Accessibility - RSA
Column 6: Absolute Surface Accessibility
Column 7: Not used
Column 8: Probability for Alpha-Helix
Column 9: Probability for Beta-strand
Column 10: Probability for Coil

E	M	NFASC_re	1	0.890999	199.5837	0	2.52E-05	8.46E-05	0.99989
E	A	NFASC_re	2	0.822632	106.1195	0	0.000178	0.000218	0.999604
E	R	NFASC_re	3	0.789607	216.3524	0	0.000201	0.000312	0.999488
E	Q	NFASC_re	4	0.780665	175.6495	0	5.27E-05	0.000324	0.999623
E	P	NFASC_re	5	0.716381	113.9045	0	2.86E-05	0.00016	0.999811
E	P	NFASC_re	6	0.703382	111.8378	0	3.21E-05	0.000161	0.999807
E	P	NFASC_re	7	0.6822	108.4698	0	7.56E-05	0.000202	0.999722
E	P	NFASC_re	8	0.68275	108.5572	0	0.000286	0.000226	0.999488
E	W	NFASC_re	9	0.639935	182.3814	0	0.000634	0.000545	0.998821
E	V	NFASC_re	10	0.579066	100.7574	0	0.000555	0.00063	0.998815
E	H	NFASC_re	11	0.641862	143.7771	0	0.000518	0.000702	0.99878
E	A	NFASC_re	12	0.601697	77.61896	0	0.00058	0.000626	0.998793
E	A	NFASC_re	13	0.548697	70.7819	0	0.000435	0.001032	0.998533
E	F	NFASC_re	14	0.536592	128.7821	0	0.000239	0.002572	0.99719
E	L	NFASC_re	15	0.566734	113.9135	0	0.000126	0.003813	0.996061
E	L	NFASC_re	16	0.573608	115.2952	0	7.97E-05	0.003012	0.996908
E	C	NFASC_re	17	0.54777	91.47765	0	0.000106	0.003483	0.996411
E	L	NFASC_re	18	0.644464	129.5373	0	9.82E-05	0.003267	0.996635
E	L	NFASC_re	19	0.66675	134.0167	0	8.86E-05	0.002037	0.997875
E	S	NFASC_re	20	0.747813	115.911	0	0.000147	0.001159	0.998693
E	L	NFASC_re	21	0.736005	147.937	0	0.000195	0.000498	0.999306
E	G	NFASC_re	22	0.797154	82.904	0	0.000467	0.000326	0.999207
E	G	NFASC_re	23	0.817293	84.99846	0	0.001605	0.000448	0.997946
E	A	NFASC_re	24	0.804079	103.7262	0	0.002641	0.0008	0.996559

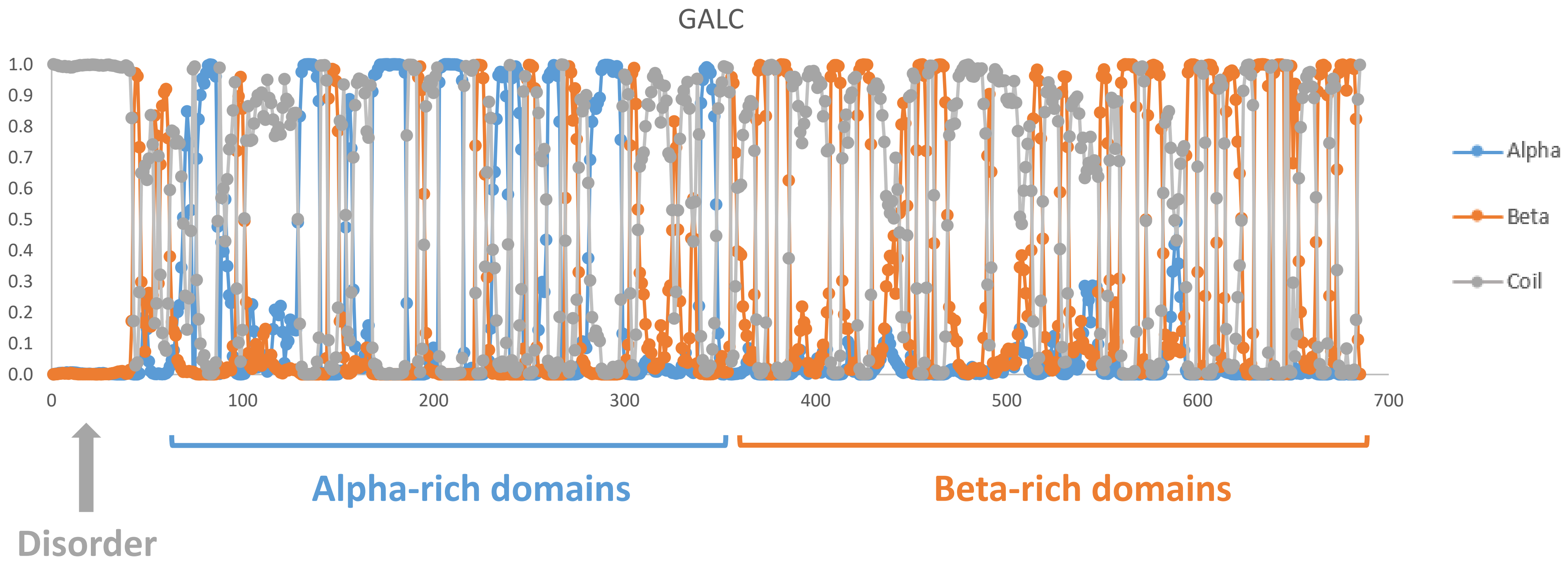
Probability of

Alpha

Beta

Coil

Secondary structure and disorder



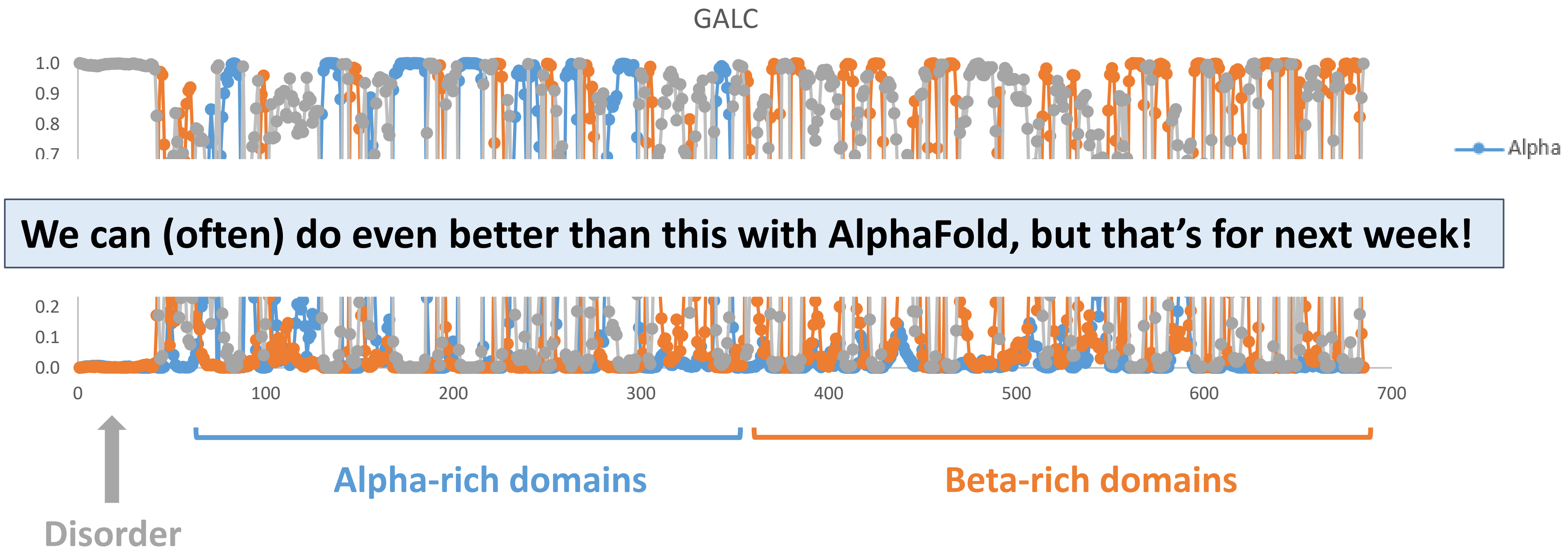
Alpha-rich domains

Beta-rich domains

Disorder

Image: Janet Deane (CC BY 4.0)

Secondary structure and disorder



We can (often) do even better than this with AlphaFold, but that's for next week!

Alpha-rich domains

Beta-rich domains

Disorder

Summary

- Today you learnt how to use online resources to:
  - Predict domains
  - Identify post translational modifications
  - Calculate the molecular mass
  - Determine the isoelectric point
  - How to calculate extinction co-efficients
  - Identify distant homologs

- Try this out with your favourite protein!

UniProt

Expasy

ProtParam tool

ApE
A plasmid Editor