FIND A GENE PROJECT

BIMM 143

Spring, 2024

Lilith Sadil
A16470107

Question 1

Protein of Interest

Protein Name & Function

The initial protein of interest chosen for this project was **lactase**.

Species & Accession Number

An NCBI search returned the amino acid sequence for a lactase protein found in ***Nothobranchius furzeri***, (accession: **KAF7215059**)

Protein Function

The protein being studied is a **glycosyl hydrolase** which specifically is involved in the cleavage of O-glycosyl compounds.

Protein Sequence

The lactase found through NCBI posesses the following amino acid sequence:

```
>KAF7215059.1 lactase [Nothobranchius furzeri]

MRTLVWVVCLLCMSGVQGGAVKDDEDFMLLAGPLMEQEDGGSACHGNDVFDYSSPMPPGSEQYFEYLQSK
GVTHFKVPLSWAQLLPTRVPSKPHQAAVSCYRTLMKRLLEVGLQPLVILHGSSVPESLRSKYGGWESQKL
VEVFQQYAEFVFREFAPSAQAWVTFSDLDEVLADGQDGDTSTLLQNILQLNKNIHQIYRLHFPGKGRLSF
GLKAKNVEKLFDFEASTLVDFLSVSIEYNCGSSTNFRQELKRLQMSSGNLPILIYKMNVHDCTDNQEQPL
GDLLQITQSRALNITGCDMKNLLGHLDMHDAPLRLDPHLTKMTRFRSSYENIWERFGAQVRADRNLFLND
SFPSGFQWATSSESFKVEGGWTEGGKGETIWDRFGHKGQAFANQTADLACDSYNKIDYDVYLLRGLQVNT
YQFSISWARIFPSGYSDSLSRKGIFYYEKLIDALVDSGIQPVATLYHWDLPQALQDNGGWTNSSIVEAFK
KYADFCFGRFGNKVKTWNTFSSPWVVSHAGYGTGEHPPGVKDHAASSYQATHNILKSHAEAWHIYNDNYR
GTYGGKVGIALNSDWAQPKDPSKPEDVAAADRYLQFMLGWFAHPIFVDGDYPAELKRQIELKKKDCPVPT
PAELPTFTPEERKRILGTSDFFGLNHYTSRLVSNNKGGCKPGPDRVGDFQAHVDPSWSSTASDWIFSAPQ
GLRLLLNYISAEYLNNTKVPIYITGNGMPTEYSADTLNDTSRVQYLHGYINEALKAIVHDGVDVQRFTVQ
SFMDGFEGRHGYSQRFGLHYVNFEDASRPRTPKQSAYFFSQIVEQNGFVAHINEVVKGLKTMRDHKSTML
SPSEVPSTSKVVWEKFSLQSKFQRQMYHHGVFPKDFSWGVSSSAYQIEGGWDADGKGPSVWDTFTQKPGS
IPDNANANVACDSYNRLDEDLYMLRGLKVKSYRFSLSWSRIFPDGRRTSLNQKGVDYYNKLIDGLLAFNI
TPMVTLHHWDLPQALQNFGGWEKTEMIDIFNDFCDFCFATFGNRVKFWMTFNQPQKIAWLGYGLGQIPPN
IKNPGIAPYKVAHKIIKAHAKAYHTYDDKYRKSQGGLISIALNADWIEPKDVNVPREVVAADRALQFQLG
WFAHPIFKNGDYPEAMKLQVGVKSDLQGLSESRLPSFTEEEKKFVRGTADMFCMNHVTTKIASHVTGRLS
PASYQYDSDFSEAEESDSPATAIQNQRAVAWGLRRILNWIKEEYGDPEIYITDNEVATGAKTTWDDIDRV
FFLKTYIDEALKAYELDGVRVKGYMTSLMDSFEWLNGYKVGFGLHHVDFANPNRPRTPKYSAHFYHQVMK
DNGFPITEDERMIHGRFRRGFMWSCATASYQIEGGWRADGKGLSIWDKFAHTPLKVFNDDNGDVACDSYN
KIDEDVAIIKQLKLTHYRFSISWSRVLPDGTTNHINEAGLGYYDRLVKALLAANVKPHITLFHWDLPQAL
QDIGGWENDTIIDKFRDYADLIFSRLGNNVKHWITINEPYNIAMVGHGYGVAAPGISFRPGTLPYIVGHN
LIKAHAEVWHLYNDKYRAVQKGIISITINSDWSEPRNPHKQEDIDAARRVVQFYIGWFAHPIYNGDYNDV
MKTVIRERSLAAGLSKSRLPEFTPEEIKRIKGTYDYFGFNHYTTVLAFPVDFGNLQHYDADRGAGTVVDR
TWLDSGSAWLKVTPFGFRRILNFIKEEYGHPPIIITENGISERGPTDLNDTLRTYYYEKYINQALKAYLI
DDVDIRGYTAWTLMDNLEWATGFSERFGLFYVNRSDSKLPRVAKASVSAFSTIIGCNGFPDPAAGPDDCL
KPRPEGTIAPITESPVMFLGMRLSTEGAETGLNATFALLIIAVIAAICLSVGFCLIKKR
```

Question 2

TBLASTN Search

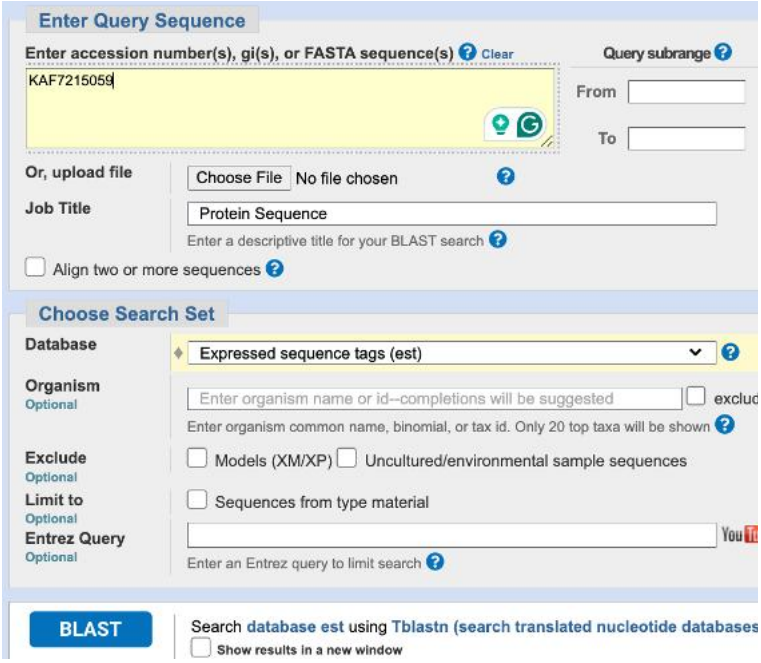
After obtaining an amino acid sequence of the ***Nothobranchius furzeri*** lactase through NCBI, a tblastn search was performed to compare the protein against existing nucleotide sequences in the database

TBLASTN Set Up

Method:
TBLASTN search against all taxa

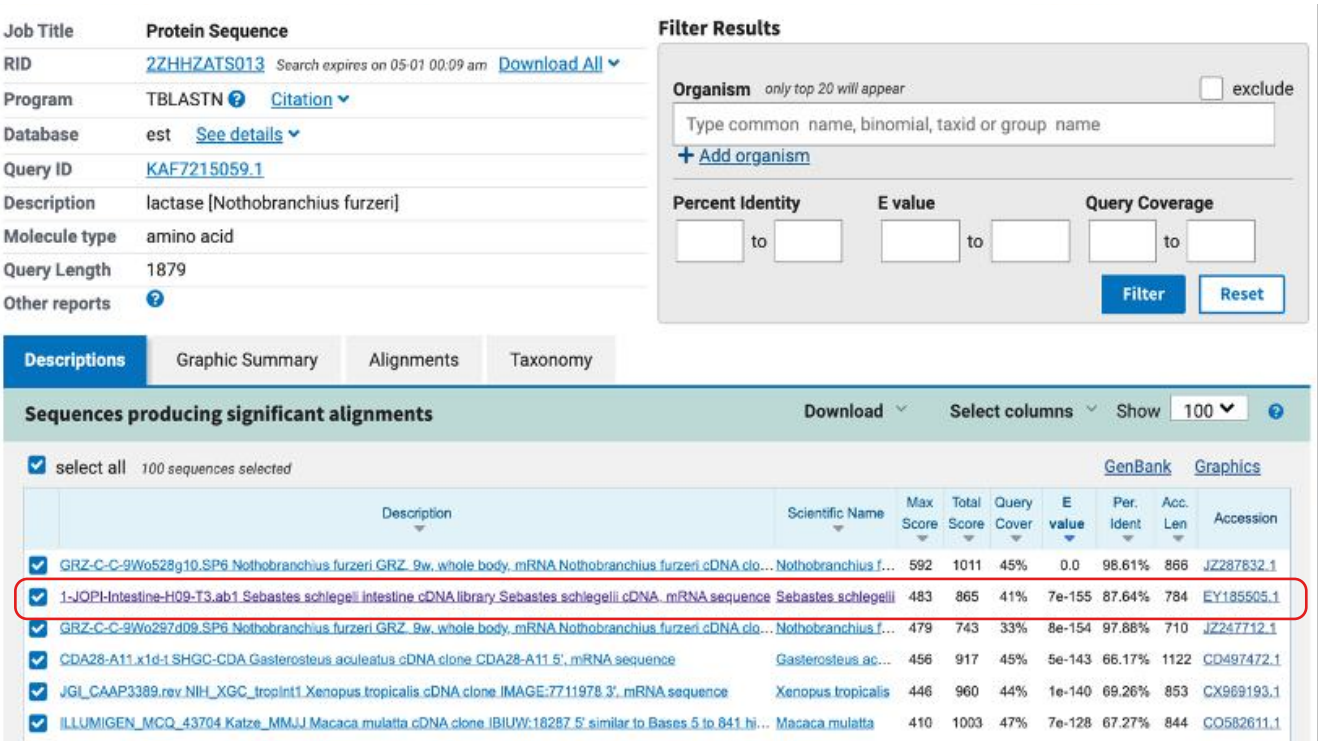
Database Searched:
Expressed Sequence Tags (est)

Limits Applied:
No taxid restrictions



TBLASTN Search Output List

Of the outputs, a 784 base pair alignment from ***Sebastes schlegelii*** (accession EY185505.1) was chosen with an **E-value** of 7e-55, a **percent identity** of 87.64%, and a **query coverage** of 41%.



Question 2

Alignment of Initial & Query Proteins:

The following diagrams depict the alignment between the starting ***Nothobranchius furzeri*** lactase sequence and the Sebastes schlegeli protein found in the BLAST database:

Download GenBank Graphics Sort by: E value ▼Next ▲Prev

1-JOPI-Intestine-H09-T3.ab1 Sebastes schlegeli intestine cDNA library Sebastes schlegelii cDNA, mRNA sequence

Sequence ID: EY185505.1 Length: 784 Number of Matches: 3

Range 1: 8 to 784 GenBank Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
483 bits(1244)	7e-155	Compositional matrix adjust.	227/259(88%)	246/259(94%)	0/259(0%)	+2

```
Query  1489  ADLIFSRLGNNVKHWITINEPYNIAMVGHGYGVAAPGISFRPGTLPYIVGHNLIKAHAEV  1548
             ADLIFSRLG+ VK WITINEPYNIA VGHGYG AAPGISFRPGTLPYIVGHNL+K+HAEV
Sbjct  8     ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV  187

Query  1549  WHLYNDKYRAVQKGIISITINSDWSEPRNPHKQEDIDAARRVVQFYIGWFAHPIYNGDYN  1608
             WHLYND YRA QKGIISITINSDWSEPRNP+KQEDIDAARRVVQFYIGWFAHP++NGDY+
Sbjct  188   WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  367

Query  1609  DVMKTVIRERSLAAGLSKSRLPEFTPEEIKRIKGTYDYFGFNHYTTVLAFPVDFGNLQHY  1668
             ++MKT+IRERSLAAGL KSRLPEFT EEIKRIKGTYDYFGFNHYTTVLAFPV++G+LQHY
Sbjct  368   NMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY  547

Query  1669  DADRGAGTVVDRTWLDSGSAWLKVTPFGFRRILNFIKEEYGHPPIIITENGISERGPTDL  1728
             DADRGAGT+ DRTWLDSGS+WLKV+PFGFRRILNFIKEEYG+P IIITENG+SERGP DL
Sbjct  548   DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPIDL  727

Query  1729  NDTLRTYYYEKYINQALKA  1747
             ND  R+YYYEKYINQALKA
Sbjct  728   NDLHRSYYYEKYINQALKA  784
```

Descriptions | **Graphic Summary** | Alignments | Taxonomy

hover to see the title click to show alignments

Alignment Scores: < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

1 sequences selected

Distribution of the top 3 Blast Hits on 1 subject sequences

Question 3

TBLASTN Match Information

Matching Sebastes schlegelii sequence from tBlastn:

From the tBLASTn search conducted, the following protein was selected:

Protein Name: 1-JOPI-Intestine-H09-T3.ab1 Sebastes schlegeli intestine cDNA, library Sebastes schlegelii cDNA, mRNA sequence

Species: Sebastes schlegelii

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Neoteleostei; Acanthomorphata; Eupercaria; Perciformes; Scorpaenoidei; Sebastidae; Sebastinae; Sebastes.

Nucleotide Sequence:

The tBLASTn search yielded the following nucleotide sequence for the potentially novel protein:

```
>EY185505.1 1-JOPI-Intestine-H09-T3.ab1 Sebastes schlegeli intestine
cDNA library Sebastes schlegelii cDNA, mRNA sequence

GCACGAGGCTGACCTCATCTTTAGCCGTCTTGGCCACAAAGTGAAATTCTGGATCACCATTAATGAGCCGTA
CAATATAGCCAACGTGGGCCATGGCTACGGAGCAGCTGCCCCAGGGATCAGTTTCCGACCAGGCACTCTGCC
CTACATTGTGGGTCACAATCTGCTTAAATCTCACGCTGAGGTCTGGCATCTCTACAATGACACGTACCGGGC
CGAACAGAAAGGAATTATCTCCATCACCATCAACTCTGACTGGTCAGAACCTAGAAATCCCTACAAGCAGGA
GGACATTGATGCCGCAAGACGTGTGGTGCAGTTCTATATTGGCTGGTTTGCCCATCCCGTATTTAATGGAGA
CTACAGCAACATGATGAAGACAATCATTCGAGAGCGGAGCTTAGCTGCTGGTCTGCTAAAATCTCGGCTGCC
TGAGTTCACTGCTGAGGAGATCAAGAGGATTAAGGGGACCTATGATTATTTTGGGTTCAACCATTACACCAC
TGTGCTTGCTTTCCCTGTGAACTATGGAGACCTTCAGCACTATGATGCCGATAGGGGTGCAGGAACAATTGC
TGATCGTACCTGGCTGGATTCAGGCTCATCCTGGCTGAAGGTCTCACCGTTTGGATTCCGCAGGATATTAAA
CTTCATTAAGGAGGAGTATGGAAATCCATCTATTATCATCACAGAAAATGGCATGTCAGAGCGGGGGCCTAT
TGACCTAAATGATCTTCATAGGAGCTACTACTATGAAAAATACATCAACCAGGCGCTGAAAGCT
```

Translated Protein Sequence (via EMBOSS, reading frame #2):

Inputting the Sebastes schlegelii nucleotide sequence into EMBOSS Transeq outputs the following amino acid sequence (reading frame 2):

```
>EMBOSS_001_2, Sebastes schlegelii novel protein

HEADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEVWHLYNDTYRA
EQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYSNMMKTIIRERSLAAGLLKSRLP
EFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHYDADRGAGTIADRTWLDSGSSWLKVSPFGFRRILN
FIKEEYGNPSIIITENGMSERGPIDLNDLHRSYYYEKYINQALKA
```

Question 4

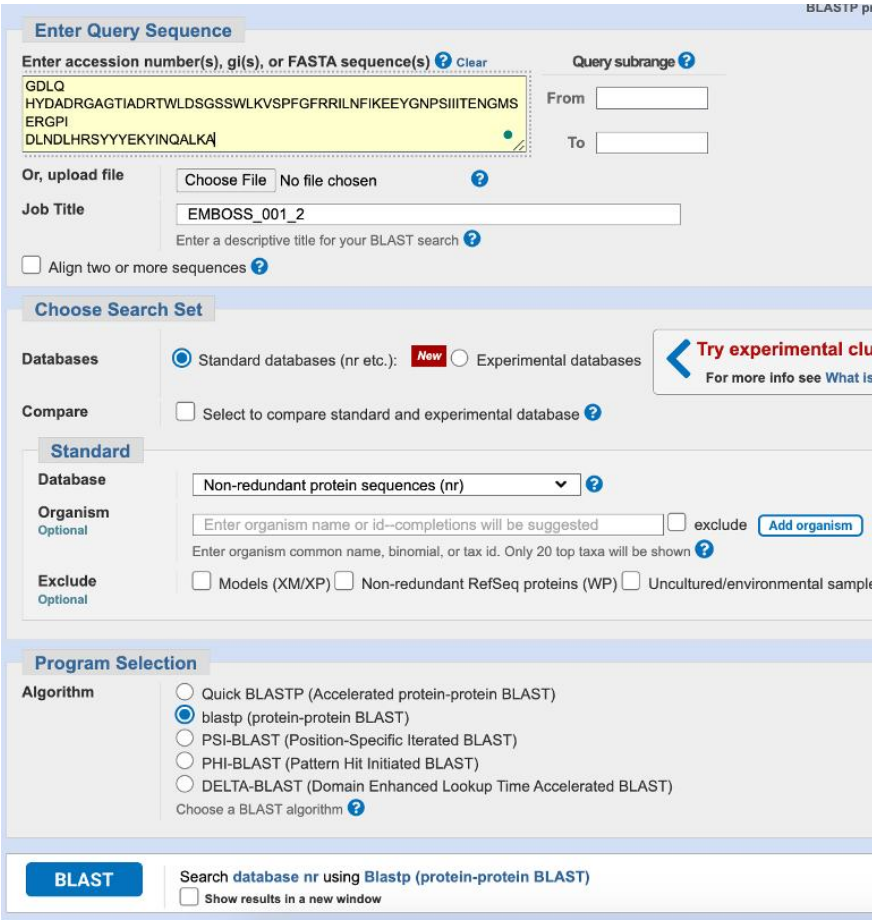
Proving the Protein is Novel

BLASTp Search

A BLASTp search of the EMBOSS-translated amino acid sequence was conducted with **no taxa restrictions** through the **non-redundant protein sequences (nr)** database:

The top results for this search have low E-values, high query coverages (~99%), and high percent identities which remain below 100%. **From this, it can be concluded that the protein found is novel**.

In other words, it can be said with confidence that an exact match to the query protein does not yet exist in the NCBI BLAST database.



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein EPR50_G00120580 [Perca flavescens]	Perca flavescens	517	517	99%	3e-180	95.75%	543	TDH07129.1
lactase-phlorizin hydrolase-like protein [Lates japonicus]	Lates japonicus	506	573	99%	1e-173	93.05%	719	GLD61503.1
lactase-phlorizin hydrolase-like [Sebastes umbrosus]	Sebastes umbrosus	532	957	99%	2e-172	98.46%	1892	XP_037646095.1
lactase/phlorizin hydrolase-like isoform X7 [Syngnathoides biaculeatus]	Syngnathoides biaculeatus	498	498	99%	8e-172	91.12%	608	XP_061697304.1
hypothetical protein PFLUV_G00142470 [Perca fluviatilis]	Perca fluviatilis	519	949	99%	2e-169	95.37%	1587	KAF1382314.1
lactase-phlorizin hydrolase [Cottoperca gobio]	Cottoperca gobio	519	948	99%	2e-169	94.98%	1572	XP_029315200.1
lactase-phlorizin hydrolase isoform X1 [Perca fluviatilis]	Perca fluviatilis	520	951	99%	3e-169	95.37%	1659	XP_039673297.1
lactase-phlorizin hydrolase-like [Sander lucioperca]	Sander lucioperca	522	956	99%	1e-168	95.75%	1888	XP_031147747.2
lactase-phlorizin hydrolase-like [Perca flavescens]	Perca flavescens	521	951	99%	1e-168	95.75%	1850	XP_028446281.1
lactase-phlorizin hydrolase [Anarrhichthys ocellatus]	Anarrhichthys ocellatus	521	954	99%	2e-168	96.14%	1895	XP_031730940.1
PREDICTED: lactase-phlorizin hydrolase-like [Stegastes partitus]	Stegastes partitus	499	721	99%	8e-168	91.51%	964	XP_008278575.1
lactase-phlorizin hydrolase-like [Kryptolebias marmoratus]	Kryptolebias marmoratus	487	487	99%	9e-168	88.80%	596	XP_017291521.3
lactase/phlorizin hydrolase-like [Anoplopoma fimbria]	Anoplopoma fimbria	519	952	99%	2e-167	94.98%	1890	XP_054471737.1
lactase/phlorizin hydrolase-like [Seriola aureovittata]	Seriola aureovittata	518	947	99%	3e-167	94.98%	1885	XP_056246129.1
lactase-phlorizin hydrolase [Etheostoma cragini]	Etheostoma cragini	518	944	99%	3e-167	94.59%	1876	XP_034741972.1
lactase-phlorizin hydrolase-like [Siniperca chuatsi]	Siniperca chuatsi	518	951	99%	3e-167	94.98%	1890	XP_044071990.1
hypothetical protein FQN60_009296 [Etheostoma spectabile]	Etheostoma spectabile	516	942	99%	5e-167	94.21%	1790	KAA8593180.1
lactase-phlorizin hydrolase-like [Seriola lalandi dorsalis]	Seriola lalandi dorsalis	517	941	99%	9e-167	94.59%	1897	XP_023278880.1
PREDICTED: lactase-phlorizin hydrolase-like [Paralichthys olivaceus]	Paralichthys olivaceus	516	936	99%	1e-166	94.59%	1884	XP_019945550.1

Question 4

Similar Alignments

The following are the three most closely aligned proteins to the novel sequence from within the database:

hypothetical protein EPR50_G00120580 [Perca flavescens]

Sequence ID: TDH07129.1 Length: 543 Number of Matches: 1

Range 1: 201 to 459 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
517 bits(1331)	3e-180	Compositional matrix adjust.	248/259(96%)	254/259(98%)	0/259(0%)

```
Query  3    ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV  62
            ADLIFSRLGHKVKFWITINEPYNIAN+GHGYGAAAPGISFRPGTLPYIVGHNLLK+HAE
Sbjct  201  ADLIFSRLGHKVKFWITINEPYNIANIGHGYGAAAPGISFRPGTLPYIVGHNLLKAHAEA  260

Query  63   WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  122
            WHLYNDTYRA+QKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS
Sbjct  261  WHLYNDTYRAKQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  320

Query  123  NMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY  182
            NMMKTIIRERSLAAGL KSRLPEFT EEIKRIKGTYDYFGFNHYTTVLAFPVNYG+LQHY
Sbjct  321  NMMKTIIRERSLAAGLSKSRLPEFTPEEIKRIKGTYDYFGFNHYTTVLAFPVNYGNLQHY  380

Query  183  DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPIDL  242
            DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIK+EYGNP IIITENGMSERGPIDL
Sbjct  381  DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKKEYGNPRIIITENGMSERGPIDL  440

Query  243  NDLHRSYYYEKYINQALKA  261
            ND+HRSYYYEKYINQ LKA
Sbjct  441  NDIHRSYYYEKYINQVLKA  459
```

Top alignment

Perca flavescens:

- Identities: 96%
- Positives: 98%
- Gaps: 0%

lactase-phlorizin hydrolase-like protein [Lates japonicus]

Sequence ID: GLD61503.1 Length: 719 Number of Matches: 2

Range 1: 320 to 578 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
506 bits(1304)	1e-173	Compositional matrix adjust.	241/259(93%)	252/259(97%)	0/259(0%)

```
Query  3    ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV  62
            ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGH+LLK+HAE
Sbjct  320  ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHHLLKAHAEA  379

Query  63   WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  122
            WHLYN+ YRA+QKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHP+FNGDYS
Sbjct  380  WHLYNEKYRAKQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPIFNGDYS  439

Query  123  NMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY  182
            +MMKTIIRERSLAAGL KSRLPEF+ EEIKRIKGTYDYFGFNHYTTVLAFPV+YG+LQHY
Sbjct  440  DMMKTIIRERSLAAGLPKSRLPEFSPEEIKRIKGTYDYFGFNHYTTVLAFPVDYGNLQHY  499

Query  183  DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPIDL  242
            DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNP IIITENG+SERG IDL
Sbjct  500  DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPPIIITENGISERGSIDL  559

Query  243  NDLHRSYYYEKYINQALKA  261
            ND+HRSYYYEKYINQ LKA
Sbjct  560  NDIHRSYYYEKYINQVLKA  578
```

Lates japonicus

- Identities: 93%
- Positives: 97%
- Gaps: 0%

lactase-phlorizin hydrolase-like [Sebastes umbrosus]

Sequence ID: XP_037646095.1 Length: 1892 Number of Matches: 3

Range 1: 1493 to 1751 GenPept Graphics ▼ Next Match ▲ Previous Mi

Score	Expect	Method	Identities	Positives	Gaps
532 bits(1371)	2e-172	Compositional matrix adjust.	255/259(98%)	255/259(98%)	0/259(0%)

```
Query  3    ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV  62
            ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAE
Sbjct  1493 ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEA  1552

Query  63   WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  122
            WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS
Sbjct  1553 WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS  1612

Query  123  NMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY  182
            NMMKTIIRERSLAAGLLKSRLPEF  EEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY
Sbjct  1613 NMMKTIIRERSLAAGLLKSRLPEFIPEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY  1672

Query  183  DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPIDL  242
            DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNP IIITENGMSERGPIDL
Sbjct  1673 DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPPIIITENGMSERGPIDL  1732

Query  243  NDLHRSYYYEKYINQALKA  261
            NDLHRSYYYEKYINQALKA
Sbjct  1733 NDLHRSYYYEKYINQALKA  1751
```

Sebastes umbrosus

- Identities: 98%
- Positives: 98%
- Gaps: 0%

Question 5

Multiple Sequence Alignment

Setting up the MSA

A multiple sequence alignment was created with the starting protein from fizcherieoii, the homologous ___ protein, and 8 additional sequences found through restricting the prior pBLAST search to the Sebastes schlegelii order (Perciformes) since no other pBLAST results were found within the Scorpaenoidei family.

1. Starting Protein: Lactase from *Nothobranchius furzeri*

```
>KAF7215059.1, Nothobranchius

ADLIFSRLGNNVKHWITINEPYNIAMVGHGYGVAAPGISFRPGTLPYIVGHNLIKAHAEV
WHLYNDKYRAVQKGIISITINSDWSEPRNPHKQEDIDAARRVVQFYIGWFAHPIYNGDYN
DVMKTVIRERSLAAGLSKSRLPEFTPEEIKRIKGTYDYFGFNHYTTVLAFPVDFGNLQHY
DADRGAGTVVDRTWLDSGSAWLKVTPFGFRRILNFIKEEYGHPPIIITENGISERGPTDL
NDTLRTYYYEKYINQALKA
```

Note: full sequence provided on page 2

2. Novel Protein: Novel lactase-homologous protein from Sebastes schlegeli

```
> Sebastes

ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV
WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVFNGDYS
NMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGDLQHY
DADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPIDL
NDLHRSYYYEKYINQALKA
```

Note: full sequence provided on page 5

2. BLAST Protein 1: lactase-phlorizin hydrolase [Cottoperca gobio]

```
XP_029315200.1, Cottoperca

ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKAHAEA
WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVFNGDYS
NMMKTIIQERSFAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTTVLAFPVNYGNLQHY
DADRGAGTIADRTWLDSGSSWLKVSPFGFRKILNFIKKEYGNPPIIITENGMSERGPIDL
NDIHRSYYYEKYINQALKA
```

Percent Match: 94.98%

3. BLAST Protein 2: PREDICTED: lactase-phlorizin hydrolase, partial [Notothenia coriiceps]

```
XP_010780417.1, Notothenia

ADLIFSRLGHKVKFWITINEPYNVANIGHGYGAAAPGISFRPGTLPYIVGHSLLKAHAEA
WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVFNGDYS
NMMKTIIRERSLAANLTTSRLPEFTPAEIKRIKGTYDYFGFNHYTTVLAFPVNYGNLQHY
DADRGTGTIADRSWLDSGSSWLKVSPFGFRRILNFIKAEYGNPPIIITENGISERGPIDL
NDTLRSYYYEQYINQLLKA
```

Percent Match: 90.73%

Question 5

4. BLAST Protein 3: hypothetical protein KUCAC02_010264 [Chaenocephalus aceratus]

KAI4805663.1, Chaenocephalus

ADLIFSRFGHKVKFWITINEPYNVANIGHGYGAAAPGISFRPGTLPYIVGHSLLKAHAEA
WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVFNGDYS
NIMKTIIRERSLADNLTKSRLPEFTPAEIKRIKGTYDYFGFNHYTTVLAFPVNYGNLQHY
DADRGAGTIADRSWLDSGSSWLKVSPFGFRRILNFIKTEYGNPPIIITENGISERGPIDL
NDTLRSYYYEQYINQLLKA

Percent Match: 89.96%

2. BLAST Protein 2: lactase-phlorizin hydrolase-like [Gasterosteus aculeatus aculeatus]

XP_040058402.1, Gasterosteus

ADLIFSRLGPKVKSWLTLNEPYNVANLGYGYGESAPGIGSRPGTLPYIVGHTLIKAHAEA
WHLYNDKYRATQKGIISISLNSDWTEPRNPYKQEDIDAARRVLQFFIGWFAHPVFNGDYS
NMMKTIIRERSLAAGLLKSRLPEFTPEEIKRIKGTHDFFGFNHYTSVLAFPVNYGDLQSY
DGDRGAGTLSDRTWLESGSFWLKVSPFGFRRVLKFIKEEYGNPPIIVTENGMSERGPIDL
NDIHRSYYYEKYINQMLKA

Percent Match: 85.33%

3. BLAST Protein 3: lactase-phlorizin hydrolase-like [Trematomus bernacchii]

XP_033990469.1, Trematomus

ADLLFSRFGRRVKFWITLNEPYIVANLGYGYGTFAPGIYNRQ----YIAAHNLIKAHAEA
WHLYNDKYRATQGGIISITINSDWAEPRNPYKQEDVDATNRYLQFFIGWFAHPIFRGDYP
ELMKTIIRERSLAVGLPESRLPEFTPEEIKRINGTHDYFGFNHYCSVLSYPVNLGKQQDY
EADRGTGTTHDRTWIESGSFWLKITPFGFRKILKFIKDEYGNPPLYVTENGISERGEVNL
NDVHRKYYYEQYINQALKA

Percent Match: 71.04%

3. BLAST Protein 4: hypothetical protein JOQ06_029923 [Pogonophryne albipinna]

KAJ4933086.1, Pogonophryne

ADLLFSRFGQRVKFWITLNEPYIVANLGYGYGTFAPGIFNRQ----YIAAHNLIKAHAEA
WHLYNDEYRATQGGIISITINSDWAEPRNPYKQEDVDATNRYLQFFIGWFAHPIFRGDYP
ELMKTIIRERSLAVGLPESGLPEFTPEEIKRINGTHDYFGLNHYCSVLSYPVNLGKQQDY
EGDRGTGVSHDRTWIESGSFWLKITPFGFRKLLKFIKDE----------
-----KYFYEQYINQALKA

Percent Match: 62.16%

3. BLAST Protein 5: Cytosolic beta-glucosidase [Dissostichus eleginoides]

KAK1890167.1, Dissostichus

AKFCFHTFGDRVKLWITINEPYVCAKLGHEDGIHAPGLK-EPGTAAYLVGHNMLRAHAMA
WHSYDSLYRAQQEGAVSLAINSDWLEPLNAGCTEDIAATERLLAFTLGWFAWPVFVTGDY
PEVMRSAIDAQSEKSG-CSSRLPSFSKDE-PAILGTADFFALNYYTSRKMKQGGGCGEKL
CMKSDRDAGEVLDPSWLICGVSWLAVVPQGLRKLLKYIKNNFNNPAVYVTENGFSQVGPL
QTEDVQRAEFYKDTIAEVAKA

Percent Match: 44.44%

Question 5

MSA Results

A CLUSTAL multiple sequence alignment by MUSCLE was produced using the 9 sequences listed prior. The results of this alignment are as follows:

```
Dissostichus        AKFCFHTFGDRVKLWITINEPYVCAKLGHEDGIHAPGLK-EPGTAAYLVGHNMLRAHAMA
Trematomus          ADLLFSRFGRRVKFWITLNEPYIVANLGYGYGTFAPGIYNRQ----YIAAHNLIKAHAEA
Pogonophryne        ADLLFSRFGQRVKFWITLNEPYIVANLGYGYGTFAPGIFNRQ----YIAAHNLIKAHAEA
Gasterosteus        ADLIFSRLGPKVKSWLTLNEPYNVANLGYGYGESAPGIGSRPGTLPYIVGHTLIKAHAEA
Nothobranchius      ADLIFSRLGNNVKHWITINEPYNIAMVGHGYGVAAPGISFRPGTLPYIVGHNLIKAHAEV
Notothenia          ADLIFSRLGHKVKFWITINEPYNVANIGHGYGAAAPGISFRPGTLPYIVGHSLLKAHAEA
Chaenocephalus      ADLIFSRFGHKVKFWITINEPYNVANIGHGYGAAAPGISFRPGTLPYIVGHSLLKAHAEA
Sebastes            ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKSHAEV
Cottoperca          ADLIFSRLGHKVKFWITINEPYNIANVGHGYGAAAPGISFRPGTLPYIVGHNLLKAHAEA

Dissostichus        WHSYDSLYRAQQEGAVSLAINSDWLEPLNAGCTEDIAATERLLAFTLGWFAWPVFVTGDY
Trematomus          WHLYNDKYRATQGGIISITINSDWAEPRNPYKQEDVDATNRYLQFFIGWFAHPIF-RGDY
Pogonophryne        WHLYNDEYRATQGGIISITINSDWAEPRNPYKQEDVDATNRYLQFFIGWFAHPIF-RGDY
Gasterosteus        WHLYNDKYRATQKGIISISLNSDWTEPRNPYKQEDIDAARRVLQFFIGWFAHPVF-NGDY
Nothobranchius      WHLYNDKYRAVQKGIISITINSDWSEPRNPHKQEDIDAARRVVQFYIGWFAHPIY-NGDY
Notothenia          WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVF-NGDY
Chaenocephalus      WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVF-NGDY
Sebastes            WHLYNDTYRAEQKGIISITINSDWSEPRNPYKQEDIDAARRVVQFYIGWFAHPVF-NGDY
Cottoperca          WHLYNDKYRAKQQGIISITINSDWSEPRNPYKQEDIDAAKRVVQFYIGWFAHPVF-NGDY

Dissostichus        PEVMRSAIDAQSEKSG-CSSRLPSFSKDE-PAILGTADFFALNYYTSRKMKQGGGCGEKL
Trematomus          PELMKTIIRERSLAVGLPESRLPEFTPEEIKRINGTHDYFGFNHYCS-VLSYPVNLGKQQ
Pogonophryne        PELMKTIIRERSLAVGLPESGLPEFTPEEIKRINGTHDYFGLNHYCS-VLSYPVNLGKQQ
Gasterosteus        SNMMKTIIRERSLAAGLLKSRLPEFTPEEIKRIKGTHDFFGFNHYTS-VLAFPVNYGDLQ
Nothobranchius      NDVMKTVIRERSLAAGLSKSRLPEFTPEEIKRIKGTYDYFGFNHYTT-VLAFPVDFGNLQ
Notothenia          SNMMKTIIRERSLAANLTTSRLPEFTPAEIKRIKGTYDYFGFNHYTT-VLAFPVNYGNLQ
Chaenocephalus      SNIMKTIIRERSLADNLTKSRLPEFTPAEIKRIKGTYDYFGFNHYTT-VLAFPVNYGNLQ
Sebastes            SNMMKTIIRERSLAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTT-VLAFPVNYGDLQ
Cottoperca          SNMMKTIIQERSFAAGLLKSRLPEFTAEEIKRIKGTYDYFGFNHYTT-VLAFPVNYGNLQ

Dissostichus        CMKSDRDAGEVLDPSWLICGVSWLAVVPQGLRKLLKYIKNNFNNPAVYVTENGFSQVGPL
Trematomus          DYEADRGTGTTHDRTWIESGSFWLKITPFGFRKILKFIKDEYGNPPLYVTENGISERGEV
Pogonophryne        DYEGDRGTGVSHDRTWIESGSFWLKITPFGFRKLLKFIKDE------------------
Gasterosteus        SYDGDRGAGTLSDRTWLESGSFWLKVSPFGFRRVLKFIKEEYGNPPIIVTENGMSERGPI
Nothobranchius      HYDADRGAGTVVDRTWLDSGSAWLKVTPFGFRRILNFIKEEYGHPPIIITENGISERGPT
Notothenia          HYDADRGTGTIADRSWLDSGSSWLKVSPFGFRRILNFIKAEYGNPPIIITENGISERGPI
Chaenocephalus      HYDADRGAGTIADRSWLDSGSSWLKVSPFGFRRILNFIKTEYGNPPIIITENGISERGPI
Sebastes            HYDADRGAGTIADRTWLDSGSSWLKVSPFGFRRILNFIKEEYGNPSIIITENGMSERGPI
Cottoperca          HYDADRGAGTIADRTWLDSGSSWLKVSPFGFRKILNFIKKEYGNPPIIITENGMSERGPI

Dissostichus        QTEDVQRAEFYKDTIAEVAKA
Trematomus          NLNDVHRKYYYEQYINQALKA
Pogonophryne        -------KYFYEQYINQALKA
Gasterosteus        DLNDIHRSYYYEKYINQMLKA
Nothobranchius      DLNDTLRTYYYEKYINQALKA
Notothenia          DLNDTLRSYYYEQYINQLLKA
Chaenocephalus      DLNDTLRSYYYEQYINQLLKA
Sebastes            DLNDLHRSYYYEKYINQALKA
Cottoperca          DLNDIHRSYYYEKYINQALKA
```

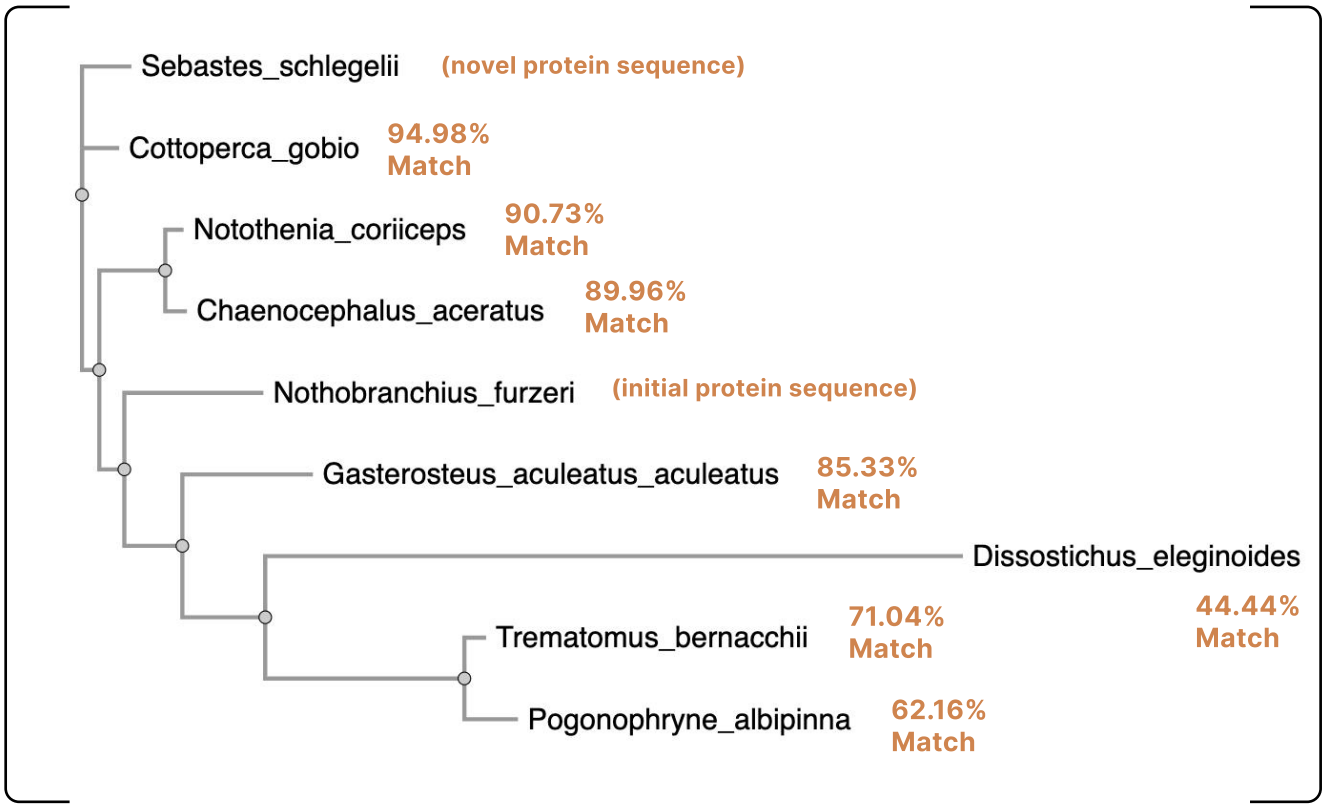
Question 6

Making a Phylogeny

MSA Results

A phylogram of the MSA results were produced via the online “Simple Phylogeny” program provided by the EBI:

MSA Phylogram



The percent match score of each sequence against the novel protein are included in orange above. As expected, the **Dissostichus eleginoides** outgroup had the lowest match score, placing it at a large horizontal distance from the other datapoints.

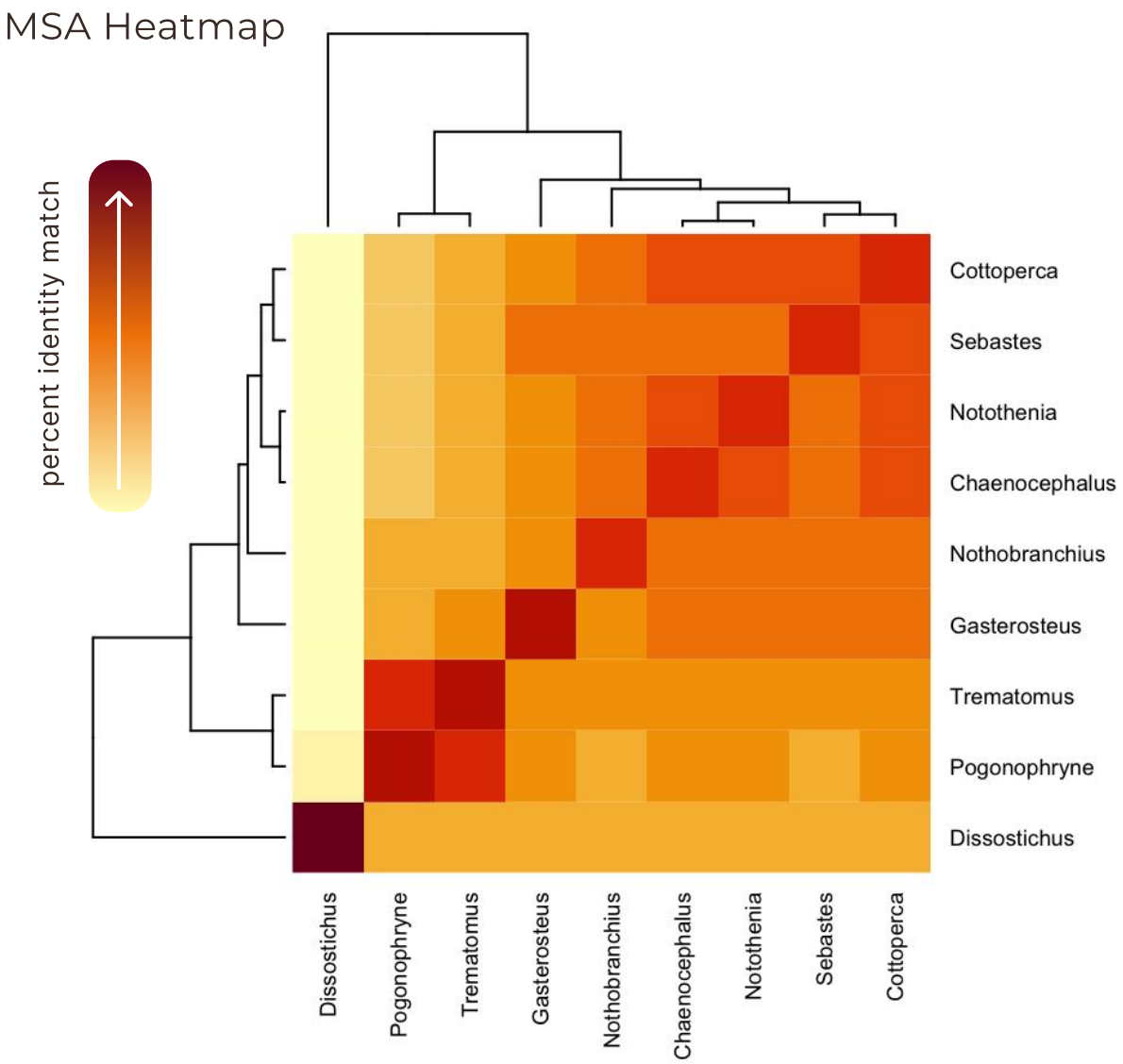
Question 7

MSA Heatmap

MSA Results

A heatmap was generated from MSA data (“result (1).fasta) using the following R code:

```
library(bio3d)
mat = bio3d::seqidentity(bio3d::read.fasta("result (1).fasta"))
matrix=as.matrix(mat)
heatmap(matrix, cexCol = 0.7, cexRow = 0.7)
```



Note: organisms’ names are referred to by their (family?) in the figure

Question 8

Finding Related Structures

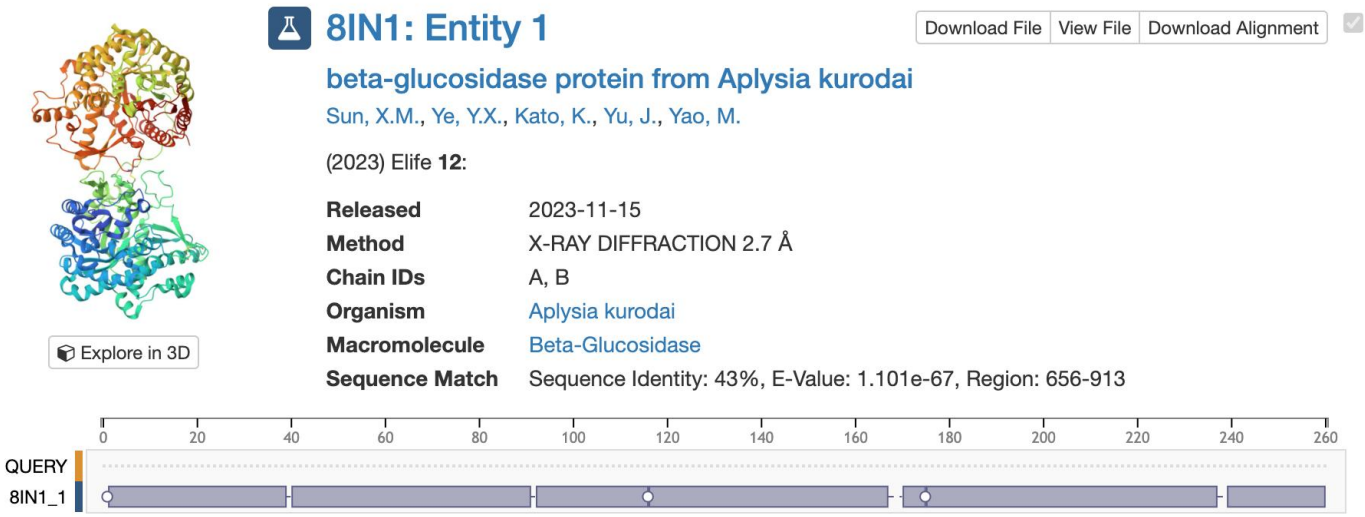
Consensus

A consensus sequence of the 9 proteins used in MSA was reconstructed using the consensus() operation in R. Since the produced alignment has many gaps, finding similar structures in existing protein databases is less reliable. Instead, the sequence for the novel protein will be searched instead.

Consensus Sequence

```
ADLIFSR-G--VKFWITINEPYN-AN-GHGYG--APGI--RPGTLPYIVGHNL-KAHAEA
WHLYNDKYRA-Q-GIISITINSDW-EPRNPYKQEDIDAA-RV-QF-IGWFAHPVF-NGDY
---MKTIIRERSLA-GL--SRLPEFTPEEIKRIKGT-DYFGFNHYT--VLAFPVN-G-LQ
-YDADRGAGT--DRTWL-SGS-WLKV-PFGFR-IL-FIK-EYGNPPII-TENG-SERGP-
DLND--R-YYYE-YINQ-LKA
```

From the RCSB PDB database, the top 3 unique protein structures with sequence similarity to the novel protein sequence are as follows:



E-Value: 1.101e-67

Sequence Similarity: 43%

Structure ID: 8IN1 (sequence: A0A1V1FXL2)

Experimental Technique: X-Ray Diffraction

Resolution: 2.70 Å

Source Organism: Aplysia kurodai

Question 8

Explore in 3D

2JFE: Entity 1

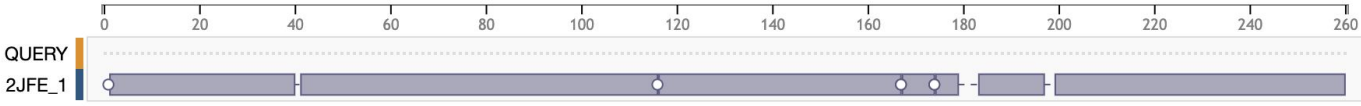
Download File | View File | Download Alignment

The crystal structure of human cytosolic beta-glucosidase

Czjzek, M., Tribolo, S., Berrin, J.G., Kroon, P.A., Juge, N.

(2007) J Mol Biol **370**: 964

Released	2007-06-19
Method	X-RAY DIFFRACTION 2.7 Å
Chain ID	A [auth X]
Organism	Homo sapiens
Macromolecule	CYTOSOLIC BETA-GLUCOSIDASE
Sequence Match	Sequence Identity: 38%, E-Value: 6.958e-52, Region: 146-403



E-Value: 6.958e-52

Sequence Similarity: 38%

Structure ID: 2JFE

Experimental Technique: X-Ray Diffraction

Resolution: 2.70 Å

Source Organism: Homo sapiens

Explore in 3D

5W21: Entity 1

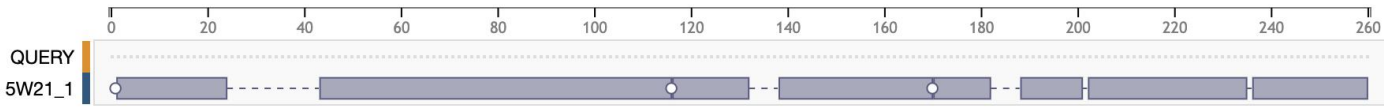
Download File | View File | Download Alignment

Crystal Structure of a 1:1:1 FGF23-FGFR1c-aKlotho Ternary Complex

Mohammadi, M.

(2018) Nature **553**: 461-466

Released	2018-01-24
Method	X-RAY DIFFRACTION 3 Å
Chain ID	A
Organism	Homo sapiens
Macromolecule	Klotho
Sequence Match	Sequence Identity: 39%, E-Value: 2.22e-53, Region: 670-901



E-Value: 2.22e-53

Sequence Similarity: 39%

Structure ID: 5W21

Experimental Technique: X-Ray Diffraction

Resolution: 3.00 Å

Source Organism: Homo sapiens

Question 9

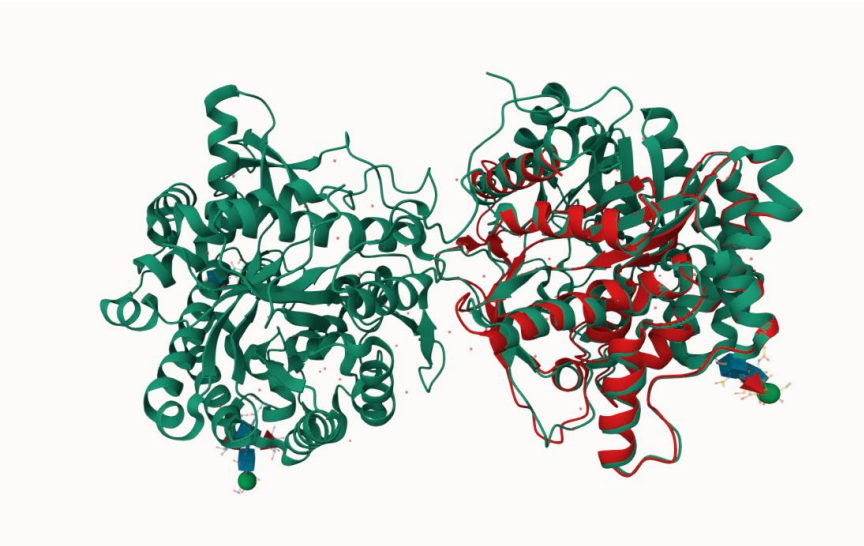
Finding Related Structures, pt. II

Comparing 3D Protein Structures

Using Mol*, a 3D rendering of the 8IN1 protein was compared against the novel protein's structure. A file of the 8IN1 3D model was accessible via the Mol* database but the novel protein had to be run through AlphaFold to first generate a structure which could be uploaded to Mol*.

A 3D rendering of the 8IN1 protein, alone

A 3D rendering of the AlphaFold-generated novel protein structures, overlayed



An overlay of the 8IN1 protein & the novel protein

Although the RCSB PDB search determined that these two sequences only have 43% sequence similarity, I believe that is because the novel protein sequence covers less than half of the 8IN1 sequence in the database. Observing these two proteins, overlayed, reveals that the structures have near-identical regions. These are likely conserved bewteen structures since both are assumed to play a role in glycosidic bond cleavage.

Question 10

ChEMBEL "Target" Search

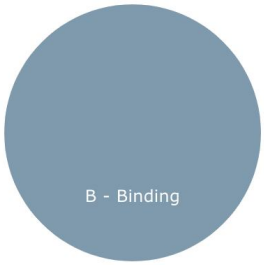
Comparing 3D Protein Structures

Looking up the novel protein sequence in ChEMBL produces the following results:

#	E-Value	Positives %	Identities %	ChEMBL ID	Name	UniProt Accessions	Organism
1.	2.7e-121	82	68.6	CHEMBL1075131	Lactase-phlorizin hydrolase	P09848	Homo sapiens
2.	3.8e-113	77.8	65.5	CHEMBL3389	Lactase-glycosylceramidase	Q02401	Rattus norvegicus
3.	6.8e-53	54.8	37.9	CHEMBL3865	Beta-glucosidase cytosolic	Q9H227	Homo sapiens
4.	4e-52	55.5	39.2	CHEMBL4523485	Klotho	Q9UEF7	Homo sapiens
5.	2.8e-47	53.1	37.7	CHEMBL4804255	FGFR4-KLB	P22455, Q86Z14	Homo sapiens
6.	2.8e-47	53.1	37.7	CHEMBL4630720	FGF1/KLB	P05230, Q86Z14	Homo sapiens
7.	2.6e-46	52.8	36.9	CHEMBL4192	Beta-glucosidase	Q43014	Prunus avium
8.	5.8e-37	46.2	34.9	CHEMBL1075057	Strictosidine beta-glucosidase	Q9M7N7	Catharanthus roseus

Associated Assays

Assays for Target CHEMBL1075131 (Lactase-phlorizin hydrolase)



B - Binding
Count: 2 (100.00%)
Click to see the data related to this category.

B - Binding

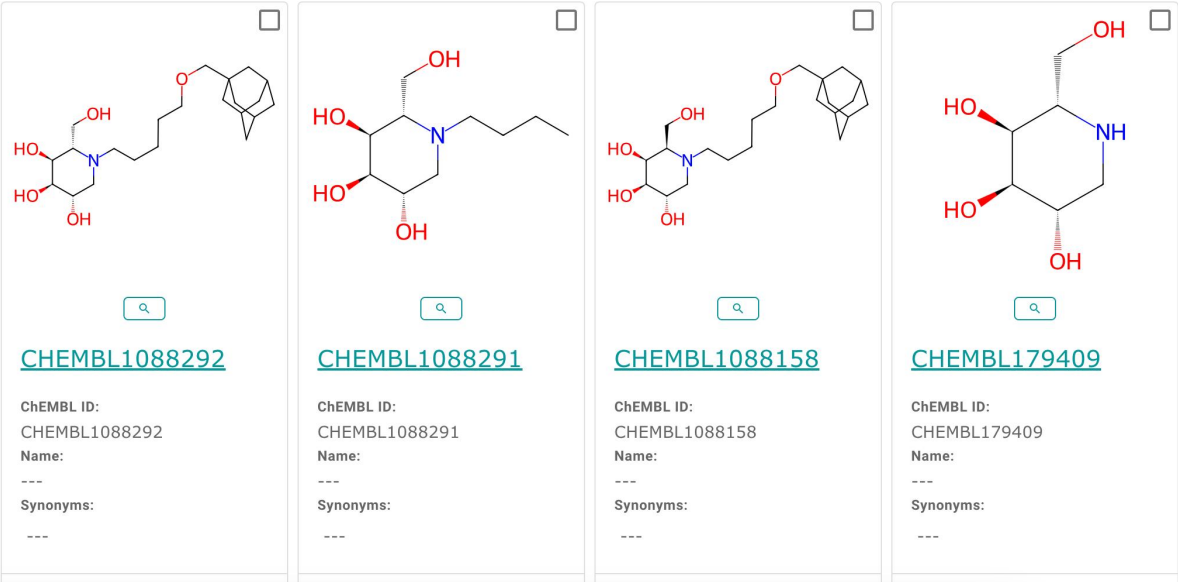
The top result, CHEMBL1075131, is a type of lactase ("lactase-phlorizin hydrolase") which has binding data for 2 different inhibitors

Question 10

The following chemical structures are the two identified inhibitors which bind to the CHEMBL1075131 lactase.

☐	#	ChEMBL ID	Assay Type	Description	Organism	Compounds	Document ChEMBL ID	BAO Format	Source
☐	1.	CHEMBL1099882	B	Inhibition of lactase by HPLC	---	13	CHEMBL1154826	single protein format	Scientific Literature
☐	2.	CHEMBL3384693	B	Inhibition of intestinal lactase (unknown origin)	Homo sapiens	24	CHEMBL3352260	single protein format	Scientific Literature

Looking into the chemical structure for the top CHEMBL1075131 lactase inhibitor, the four related structures are displayed.



Given that these compounds act as inhibitors for the CHEMBL1075131 lactase, there is a chance that they act similarity upon the novel protein being studied. Since the novel structure was obtained as a homolog to the ***Nothobranchius furzeri*** lactase enzyme, it is understandable that the novel protein also functions like a lactase and, therefore, may be inhibited in similar ways to other lactases.