Find a Gene Project

Jaewon Kim

PID: A16675101

Email: jak001@ucsd.edu

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: major intrinsically disordered NOTCH2-binding receptor 1-like (MINAR2)
Accession: NP_001244237.1
Species: Homo Sapiens
Function: MINAR2 prevents cancer development and activate cell defense system by negatively regulating following:

- Angiogenesis
- Cell growth
- Cell population proliferation
- Neuron projection development
- Protein ubiquitination
- TOR signaling

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Blast method: tblastn
Database searched: Expressed sequence tags (est)
Limited organism: Humans (9606)

Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is not necessary to print out all of the blast results if there are many pages.

Chosen match- top match (highlighted with red box in the screenshot above)
Accession: DR731264.1
Name: MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone MGC7.2.1.1.1.H04, mRNA sequence
Species: Homo sapiens
Total/Max score: 147, 147
Query cover, E-value, percent identity: 41%, 1e-44, 92.41% (other scores in screenshot above and below)

MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone MGC7.2.1.1.1.H04, mRNA sequence

Sequence ID: DR731264.1 Length: 239 Number of Matches: 1

Range 1: 3 to 239 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
147 bits(372)	1e-44	Compositional matrix adjust.	73/79(92%)	74/79(93%)	0/79(0%)	-1

```
Query  28    ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV  87
             +L QASLVRFPGGNYPAA HWQNLVYSQREKKNIAAQRIR  SADSLVTADSPPPSMSSV
Sbjct  239   SLSQASLVRFPGGNYPAA*HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSV  60

Query  88    MKNNPLYGDLSLEEAMEER  106
             MKNNPLY DLSLEEAMEER
Sbjct  59    MKNNPLYSDLSLEEAMEER  3
```

Related Information
Gene - associated gene details

Alignment details:

Query: major intrinsically disordered NOTCH2-binding receptor 1-like [Homo sapiens]
Query ID: NP_001244237.1 Length: 190

>MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone MGC7.2.1.1.1.H04, mRNA sequence
Sequence ID: DR731264.1 Length: 239

Range 1: 3 to 239

Score:147 bits(372), Expect:1e-44, Method: Compositional matrix adjust.
Identities:73/79(92%), Positives:74/79(93%), Gaps:0/79(0%)

Query 28   ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV 87
            +L QASLVRFPGGNYPAA   HWQNLVYSQREKKNIAAQRIR   SADSLVTADSPPPSMSSV
Subject 239 SLSQASLVRFPGGNYPAA*HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSV 60

Query 88   MKNNPLYGDLSLEEAMEER 106
            MKNNPLY DLSLEEAMEER
Subject 59 MKNNPLYSDLSLEEAMEER 3

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

In general, [Q2] is the most difficult for students because it requires you to have a "feel" for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not "novel"), a near match (something that might be "novel", depending on the results of [Q4]), and a non-homologous result. If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[Q3] Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

EMBOSS Transeq result:

>DR731264.1_1 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

FLSSIASSKLRSL*SGLFFITDDMDGGGLSAVTRLSALKNLIR*AAIFFFSLCE*TRFCQ
CYAAG*FPPGNLTRLA*ERX

>DR731264.1_2 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

FFLP*LPPNLGHYRVGYSS*LMTWMVGGYQQ*QGCLH*KT*FVEQQYSSFPSVSRQGFAS
VMQQDNFHPETSPGWPERGX

>DR731264.1_3 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

SFFHSFLQT*VTIEWVILHN**HGWWGAISSDKAVCTEKPNSLSSNILLFPL*VDKVLPV
LCSRIISTRKPHQAGLREG

>DR731264.1_4 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

PLSGQPGEVSGWKLSCCITLAKPCLLTEGKEEYCCSTN*VFQCRQPCHC**PPTIHVISY
EE*PTL**PKFGGSYGRKK

>DR731264.1_5 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

PSLRPAW*GFRVEIILLHNTGKTLSTHRGKRRILLLNELGFSVQTALSLLIAPHHPCHQL
*RITHSIVT*VWRKLWKKEX

>DR731264.1_6 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Homo sapiens cDNA clone
MGC7.2.1.1.1.H04, mRNA sequence

SLSQASLVRFPGGNYPAA*HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSV
MKNNPLYSDLSLEEAMEERX

Chosen sequence for blastp was highlighted in red. Green is sequences that returned no result when blastp was done.

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as S. cerevisiae, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Protein name: Homo sapiens major intrinsically disordered NOTCH2-binding receptor (MINAR)

Species: Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

 • If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.

• If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.

• If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.

• If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

None of blastp result had 100% identity, and <95% identity with low E-value indicates that the results are likely to be a novel protein. The top match was "major intrinsically disordered NOTCH2-binding receptor 1-like" from Homo sapiens, where there total score/max score/query coverage/e-value/percent identity are 148/148/96%/5e-32/92.41, respectively. Alignment detail is shown in the screenshot below.

*Result wasn't filtered just for Homo sapiens as top hit was the only result from corresponding species.

Job Title	DR731264.1_6 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331...
RID	X6F3P1SZ016 *Search expires on 02-20 18:33 pm* Download All ▾
Program	BLASTP ❓ Citation ▾
Database	nr See details ▾
Query ID	lcl|Query_11709707
Description	DR731264.1_6 MGC7.2.1.1.1.H04.R.1 NIH_MGC_331 Hom ...
Molecule type	amino acid
Query Length	80
Other reports	Distance tree of results Multiple alignment MSA viewer ❓

Filter Results

Organism *only top 20 will appear*   ☐ exclude

[Type common name, binomial, taxid or group name]

➕ Add organism

Percent Identity	E value	Query Coverage
[ ] to [ ]	[ ] to [ ]	[ ] to [ ]

                                          Filter   **Reset**

Compare these results against the new Clustered nr database ❓   **BLAST**   ✕

Descriptions	Graphic Summary	Alignments	Taxonomy

Sequences producing significant alignments

Download ▾   Select columns ▾   Show [100 ▾] ❓

☑ select all *100 sequences selected*   GenPept   Graphics   Distance tree of results   Multiple alignment   MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Homo sapiens]	Homo sapiens	148	148	96%	5e-43	92.41%	190	NP_001244237.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Pan paniscus]	Pan paniscus	147	147	96%	2e-42	91.14%	190	XP_003829382.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Pan troglodytes]	Pan troglodytes	147	147	96%	2e-42	91.14%	190	XP_003950663.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Pongo abelii]	Pongo abelii	146	146	96%	3e-42	91.14%	190	XP_002815898.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Rhinopithecus roxellana]	Rhinopithecus roxellana	145	145	96%	5e-42	91.14%	190	XP_010379869.2
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	145	145	96%	9e-42	89.87%	190	XP_004042492.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Nomascus leucogenys]	Nomascus leucogenys	145	145	95%	1e-41	93.42%	190	XP_004087485.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Piliocolobus tephrosceles]	Piliocolobus tephrosceles	144	144	96%	2e-41	89.87%	190	XP_023050104.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Trachypithecus francoisi]	Trachypithecus francoisi	144	144	96%	2e-41	89.87%	190	XP_033037245.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog isoform X3 [Cercocebus atys]	Cercocebus atys	144	144	96%	2e-41	89.87%	190	XP_011945527.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog [Mandrillus leucophaeus]	Mandrillus leucophaeus	144	144	96%	3e-41	89.87%	190	XP_011836578.1
☑	UPF0258 protein KIAA1024-like homolog isoform X2 [Macaca nemestrina]	Macaca nemestrina	142	142	96%	6e-41	87.34%	162	XP_024644098.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog [Colobus angolensis palliatus]	Colobus angolensis pal...	143	143	96%	7e-41	88.61%	190	XP_011797493.1
☑	UPF0258 protein KIAA1024-like homolog [Theropithecus gelada]	Theropithecus gelada	142	142	96%	9e-41	88.61%	190	XP_025245224.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Chlorocebus sabaeus]	Chlorocebus sabaeus	142	142	96%	1e-40	88.61%	190	XP_008012506.2
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Macaca mulatta]	Macaca mulatta	142	142	96%	1e-40	87.34%	190	NP_001180990.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Symphalangus syndactylus]	Symphalangus syndact...	142	142	95%	1e-40	92.11%	190	XP_055155262.1
☑	UPF0258 protein KIAA1024-like homolog isoform X1 [Macaca nemestrina]	Macaca nemestrina	142	142	96%	1e-40	87.34%	190	XP_011715000.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Hylobates moloch]	Hylobates moloch	142	142	95%	2e-40	90.79%	190	XP_032012416.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog [Rhinopithecus bieti]	Rhinopithecus bieti	141	141	96%	4e-40	88.61%	190	XP_017730223.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Sapajus apella]	Sapajus apella	139	139	96%	2e-39	86.08%	190	XP_032117284.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Callithrix jacchus]	Callithrix jacchus	137	137	96%	9e-39	84.81%	190	XP_002744678.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Cebus imitator]	Cebus imitator	133	133	96%	4e-37	83.54%	190	XP_017387650.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog isoform X2 [Cercocebus atys]	Cercocebus atys	133	133	96%	5e-37	72.45%	193	XP_011945526.1
☑	UPF0258 protein [Sciurus carolinensis]	Sciurus carolinensis	133	133	96%	6e-37	82.28%	194	MBZ3891378.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Sciurus carolinensis]	Sciurus carolinensis	133	133	96%	6e-37	82.28%	194	XP_047412257.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Saimiri boliviensis boliviensis]	Saimiri boliviensis boliv...	132	132	96%	8e-37	82.28%	190	XP_003920680.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Mesoplodon densirostris]	Mesoplodon densirostris	132	132	96%	1e-36	78.48%	190	XP_059948605.1
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Physeter catodon]	Physeter catodon	132	132	96%	1e-36	78.48%	190	XP_007122371.2
☑	major intrinsically disordered NOTCH2-binding receptor 1-like [Balaenoptera musculus]	Balaenoptera musculus	132	132	96%	2e-36	78.48%	190	XP_036701798.1
☑	PREDICTED: UPF0258 protein KIAA1024-like homolog isoform X1 [Cercocebus atys]	Cercocebus atys	132	132	96%	2e-36	72.45%	209	XP_011945525.1

major intrinsically disordered NOTCH2-binding receptor 1-like [Homo sapiens]

Sequence ID: NP_001244237.1   Length: **190**   Number of Matches: **1**

See 4 more title(s) ⌄   See all Identical Proteins(IPG)

Range 1: 28 to 106 GenPept   Graphics   ▼ Next Match   ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
148 bits(374)	5e-43	Compositional matrix adjust.	73/79(92%)	74/79(93%)	0/79(0%)

```
Query  1    SLSQASLVRFPGGNYPAA*HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSV  60
            +L QASLVRFPGGNYPAA HWQNLVYSQREKKNIAAQRIR  SADSLVTADSPPPSMSSV
Sbjct  28   ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV  87

Query  61   MKNNPLYSDLSLEEAMEER  79
            MKNNPLY DLSLEEAMEER
Sbjct  88   MKNNPLYGDLSLEEAMEER  106
```

5. Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

Multiple sequence alignment (MSA) was done with following sequences:
- Original query protein (red)
- Novel protein (red)
- 18 top match proteins from BLASTP result in Q.4 (blue)

To display species names in MSA, name of species were added in front of the protein names. Since all proteins are MINAR2, only species name was shown. For Macaca nemestrina, X1 and X2 were added at the end of the name to differentiate X1 and X2 isoforms (same protein/species name, different sequence).

Relabeled sequences for alignment:

```
>Homo_sapiens(Original) NP_001244237.1 major intrinsically disordered NOTCH2-
binding receptor 1-like [Homo sapiens]

ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAME
ER

>Homo_sapiens(Novel)

SLSQASLVRFPGGNYPAA-
HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSVMKNNPLYSDLSLEEAMEERX
```

```
>Pan_paniscus XP_003829382.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Pan paniscus]

MDLSVLPNNNHPEKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRGSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVVISILVTIVTIITFFT

>Pan_troglodytes XP_003950663.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Pan troglodytes]

MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRRSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKHSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVVISILVTIVTMITFFT

>Pongo_abelii XP_002815898.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Pongo abelii]
```

```
MDLSVLPNNNHPDKFLQLDVKSLTRNSALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVVISILVTIVTIITFST
```

>Rhinopithecus_roxellana XP_010379869.2 major intrinsically disordered
NOTCH2-binding receptor 1-like [Rhinopithecus roxellana]

```
MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKRNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLAIISILVAIVTIITFFT
```

>Gorilla_gorilla_gorilla XP_004042492.1 major intrinsically disordered
NOTCH2-binding receptor 1-like [Gorilla gorilla gorilla]

```
MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQKEKKNIAAQRIRGSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVVISILVTIVTIITFFT
```

>Nomascus_leucogenys XP_004087485.1 major intrinsically disordered NOTCH2-
binding receptor 1-like [Nomascus leucogenys]

```
MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVIISILVIIVTIITFFT
```

>Piliocolobus_tephrosceles XP_023050104.1 major intrinsically disordered
NOTCH2-binding receptor 1-like [Piliocolobus tephrosceles]

```
MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRVGLILLAVISILVAIVTIITFFT
```

>Trachypithecus_francoisi XP_033037245.1 major intrinsically disordered
NOTCH2-binding receptor 1-like [Trachypithecus francoisi]

```
MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLAIISILVAIVTIITFFT
```

>Macaca_nemestrina_X2 XP_024644098.1 UPF0258 protein KIAA1024-like homolog
isoform X2 [Macaca nemestrina]

```
MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKEVKQSLCMPDMWQPSTRKISGSR
VMEETPDF
```

>heropithecus_gelada XP_025245224.1 UPF0258 protein KIAA1024-like homolog
[Theropithecus gelada]

```
MDLSVLPNNNHPDKFLQLDVKSFTRSSALLQTSLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLAIISILVAIVTIITFFT
```

>Chlorocebus_sabaeus XP_008012506.2 major intrinsically disordered NOTCH2-
binding receptor 1-like [Chlorocebus sabaeus]

MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTA
ESPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYEKRSRHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLAIISILVAIVTIITFFT

>Macaca_mulatta NP_001180990.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Macaca mulatta]

MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFRRMGLILLAIISILVAIVTIITFFT

>Symphalangus_syndactylus XP_055155262.1 major intrinsically disordered
NOTCH2-binding receptor 1-like [Symphalangus syndactylus]

MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEGRKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVIISILVIIVTIITFFT

>Macaca_nemestrina_X1 XP_011715000.1 UPF0258 protein KIAA1024-like homolog
isoform X1 [Macaca nemestrina]

MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNNLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLAIISILVAIVTIITFFT

>Hylobates_moloch XP_032012416.1 major intrinsically disordered NOTCH2-
binding receptor 1-like [Hylobates moloch]

MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAARRIRGSSADSLLTA
DSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKHSKFCRMGLILLVIISILVIIVTIITFFT

>Sapajus_apella XP_032117284.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Sapajus apella]

MDLSVLPNNNHPDKFLQLDVKSLMRSSALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGPSAESLVTA
DSPPPSMSSIMKNNPLYGDLSLEEAMEERKKSPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKQSKFCRMGLILLVITSILVTIVTIITFFT

>Callithrix_jacchus XP_002744678.1 major intrinsically disordered NOTCH2-
binding receptor 1-like [Callithrix jacchus]

MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARYPCGNYPAAQHWQNLVYSQREKNNIAAQRIRGSSAESLVTA
DSPPPSMSSIMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKQSKFCRMGLILLVVTSILVTIVTIITFFT

>Cebus_imitator XP_017387650.1 major intrinsically disordered NOTCH2-binding
receptor 1-like [Cebus imitator]

MDLSVLPNNNHPDKFLQLDVKSLMRSSALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGSNAASLVTA
DSPPPSMSSTMKNNPLCGDLSLEEAMEERKKSPSWTIEEYDKRSLHTNLSGHLKENPNDLRFWLGDMYTPGFDTLLK
KEEKQEKQSKFCRMGLILLVITSILVTIVTIITFFT

Before MSA, sequences were trimmed so that there is no hangovers between novel/original sequence and other blasted results (trimmed sequences below). For non-trimmed MSA, please see screenshot below the alignment result

>Homo_sapiens(Novel)
SLSQASLVRFPGGNYPAA-HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSVMKNNPLYSDLSLEEAMEERX

>Cebus_imitator
ALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGSNAASLVTADSPPPSMSSTMKNNPLCGDLSLEEAMEERK

>Callithrix_jacchus
ALLQASLARYPCGNYPAAQHWQNLVYSQREKNNIAAQRIRGSSAESLVTADSPPPSMSSIMKNNPLYGDLSLEEAMEERK

>Sapajus_apella
ALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGPSAESLVTADSPPPSMSSIMKNNPLYGDLSLEEAMEERK

>Hylobates_moloch
AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAARRIRGSSADSLLTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Macaca_nemestrina_X2
ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSV
MKNNPLYGDLSLEEAMEERK

>Macaca_mulatta
ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Macaca_nemestrina_X1
ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Chlorocebus_sabaeus
ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTAESPPPSMSSVMKNNPLYGDLSLEEAMEERK

>heropithecus_gelada
ALLQTSLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Rhinopithecus_roxellana
ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Piliocolobus_tephrosceles
ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Trachypithecus_francoisi
ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Symphalangus_syndactylus
AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEGRK

>Nomascus_leucogenys
AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Pongo_abelii
ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Gorilla_gorilla_gorilla
ALLQASLARFPGGNYPAAQHWQNLVYSQKEKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Homo_sapiens(Original)
ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEER-

>Pan_troglodytes
ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRRSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

>Pan_paniscus
ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERK

Alignment result:
Obtained using MUSCLE (ver.3.8) at EBI

CLUSTAL multiple sequence alignment by MUSCLE (3.8)


```
Homo_sapiens(Novel)         SLSQASLVRFPGGNYPAA-HWQNLVYSQREKKNIAAQRIRFFSADSLVTADSPPPSMSSV
Cebus_imitator              ALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGSNAASLVTADSPPPSMSST
Callithrix_jacchus          ALLQASLARYPCGNYPAAQHWQNLVYSQREKNNIAAQRIRGSSAESLVTADSPPPSMSSI
Sapajus_apella              ALLHASLARFPGGNYPAAQHWQNLVYSQREKNNIAAQRIRGPSAESLVTADSPPPSMSSI
Hylobates_moloch            AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAARRIRGSSADSLLTADSPPPSMSSV
Macaca_nemestrina_X2        ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSV
Macaca_mulatta              ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSV
Macaca_nemestrina_X1        ALLQASMARFPGGNYPAAQHWQNLVYSQREKKNIAGQRIRGSSAESLVTADSPPPSMSSV
Chlorocebus_sabaeus         ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSAESLVTAESPPPSMSSV
heropithecus_gelada         ALLQTSLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSV
Rhinopithecus_roxellana     ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSV
Piliocolobus_tephrosceles   ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSV
Trachypithecus_francoisi    ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSAESLVTADSPPPSMSSV
Symphalangus_syndactylus    AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV
Nomascus_leucogenys         AVLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV
Pongo_abelii                ALLQASLARFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV
Gorilla_gorilla_gorilla     ALLQASLARFPGGNYPAAQHWQNLVYSQKEKKNIAAQRIRGSSADSLVTADSPPPSMSSV
Homo_sapiens(Original)      ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIAAQRIRGSSADSLVTADSPPPSMSSV
Pan_troglodytes             ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRRSSADSLVTADSPPPSMSSV
Pan_paniscus                ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKNIATQRIRGSSADSLVTADSPPPSMSSV
                            :: ::*:.*:* ****** *********.**:*** .***   .* **:**:********

Homo_sapiens(Novel)         MKNNPLYSDLSLEEAMEERX
Cebus_imitator              MKNNPLCGDLSLEEAMEERK
Callithrix_jacchus          MKNNPLYGDLSLEEAMEERK
Sapajus_apella              MKNNPLYGDLSLEEAMEERK
Hylobates_moloch            MKNNPLYGDLSLEEAMEERK
Macaca_nemestrina_X2        MKNNPLYGDLSLEEAMEERK
Macaca_mulatta              MKNNPLYGDLSLEEAMEERK
Macaca_nemestrina_X1        MKNNPLYGDLSLEEAMEERK
Chlorocebus_sabaeus         MKNNPLYGDLSLEEAMEERK
heropithecus_gelada         MKNNPLYGDLSLEEAMEERK
Rhinopithecus_roxellana     MKNNPLYGDLSLEEAMEERK
Piliocolobus_tephrosceles   MKNNPLYGDLSLEEAMEERK
Trachypithecus_francoisi    MKNNPLYGDLSLEEAMEERK
Symphalangus_syndactylus    MKNNPLYGDLSLEEAMEGRK
Nomascus_leucogenys         MKNNPLYGDLSLEEAMEERK
Pongo_abelii                MKNNPLYGDLSLEEAMEERK
Gorilla_gorilla_gorilla     MKNNPLYGDLSLEEAMEERK
Homo_sapiens(Original)      MKNNPLYGDLSLEEAMEER-
Pan_troglodytes             MKNNPLYGDLSLEEAMEERK
Pan_paniscus                MKNNPLYGDLSLEEAMEERK
                            ****** .********* *
```

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Homo_sapiens(Novel)          --------------------------SLSQASLVRFPGGNYPAA-HWQNLVYSQREKKN
Macaca_nemestrina_X2         MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKN
Hylobates_moloch             MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Symphalangus_syndactylus     MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Nomascus_leucogenys          MDLSVLPNNNHPDKFLQLDVKSLTRSSAVLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Macaca_mulatta               MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKN
Macaca_nemestrina_X1         MDLSVLPNNNHPDKFLQLDIKSLTRSSALLQASMARFPGGNYPAAQHWQNLVYSQREKKN
Pan_troglodytes              MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKN
Gorilla_gorilla_gorilla      MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Chlorocebus_sabaeus          MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Piliocolobus_tephrosceles    MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKN
heropithecus_gelada          MDLSVLPNNNHPDKFLQLDVKSFTRSSALLQTSLARFPGGNYPAAQHWQNLVYSQREKKN
Rhinopithecus_roxellana      MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKN
Trachypithecus_francoisi     MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Pan_paniscus                 MDLSVLPNNNHPEKFLQLDVKSLTRSSALLQASLVRFPGGNYPAAQHWQNLVYSQREKKN
Pongo_abelii                 MDLSVLPNNNHPDKFLQLDVKSLTRNSALLQASLARFPGGNYPAAQHWQNLVYSQREKKN
Homo_sapiens(Original)       --------------------------ALLQASLVRFPGGNYPAAQHWQNLVYSQREKKN
Callithrix_jacchus           MDLSVLPNNNHPDKFLQLDVKSLTRSSALLQASLARYPCGNYPAAQHWQNLVYSQREKNN
Sapajus_apella               MDLSVLPNNNHPDKFLQLDVKSLMRSSALLHASLARFPGGNYPAAQHWQNLVYSQREKNN
Cebus_imitator               MDLSVLPNNNHPDKFLQLDVKSLMRSSALLHASLARFPGGNYPAAQHWQNLVYSQREKNN
                                       ::  ::*:.*:* ****** *********.**.*


Homo_sapiens(Novel)          IAAQRIRFFSADSLVTADSPPPSMSSVMKNNPLYSDLSLEEAMEERX--------------
Macaca_nemestrina_X2         IAGQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Hylobates_moloch             IAARRIRGSSADSLLTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Symphalangus_syndactylus     IAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEGRKKNPSWTIEEYDKR
Nomascus_leucogenys          IAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Macaca_mulatta               IAGQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Macaca_nemestrina_X1         IAGQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Pan_troglodytes              IATQRIRRSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKH
Gorilla_gorilla_gorilla      IAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Chlorocebus_sabaeus          IAAQRIRGSSAESLVTAESPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYEKR
Piliocolobus_tephrosceles    IAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
heropithecus_gelada          IAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Rhinopithecus_roxellana      IAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKRNPSWTIEEYDKR
Trachypithecus_francoisi     IAAQRIRGSSAESLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Pan_paniscus                 IATQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Pongo_abelii                 IAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Homo_sapiens(Original)       IAAQRIRGSSADSLVTADSPPPSMSSVMKNNPLYGDLSLEEAMEER--------------
Callithrix_jacchus           IAAQRIRGSSAESLVTADSPPPSMSSIMKNNPLYGDLSLEEAMEERKKNPSWTIEEYDKR
Sapajus_apella               IAAQRIRGPSAESLVTADSPPPSMSSIMKNNPLYGDLSLEEAMEERKKSPSWTIEEYDKR
Cebus_imitator               IAAQRIRGSNAASLVTADSPPPSMSSTMKNNPLCGDLSLEEAMEERKKSPSWTIEEYDKR
                             ** .***  .* **.**.********* ****** .********* *


Homo_sapiens(Novel)          ----------
Macaca_nemestrina_X2         ----------
Hylobates_moloch             IIVTIITFFT
Symphalangus_syndactylus     IIVTIITFFT
Nomascus_leucogenys          IIVTIITFFT
Macaca_mulatta               AIVTIITFFT
Macaca_nemestrina_X1         AIVTIITFFT
Pan_troglodytes              TIVTMITFFT
Gorilla_gorilla_gorilla      TIVTIITFFT
Chlorocebus_sabaeus          AIVTIITFFT
Piliocolobus_tephrosceles    AIVTIITFFT
heropithecus_gelada          AIVTIITFFT
Rhinopithecus_roxellana      AIVTIITFFT
Trachypithecus_francoisi     AIVTIITFFT
Pan_paniscus                 TIVTIITFFT
Pongo_abelii                 TIVTIITFST
Homo_sapiens(Original)       ----------
Callithrix_jacchus           TIVTIITFFT
```
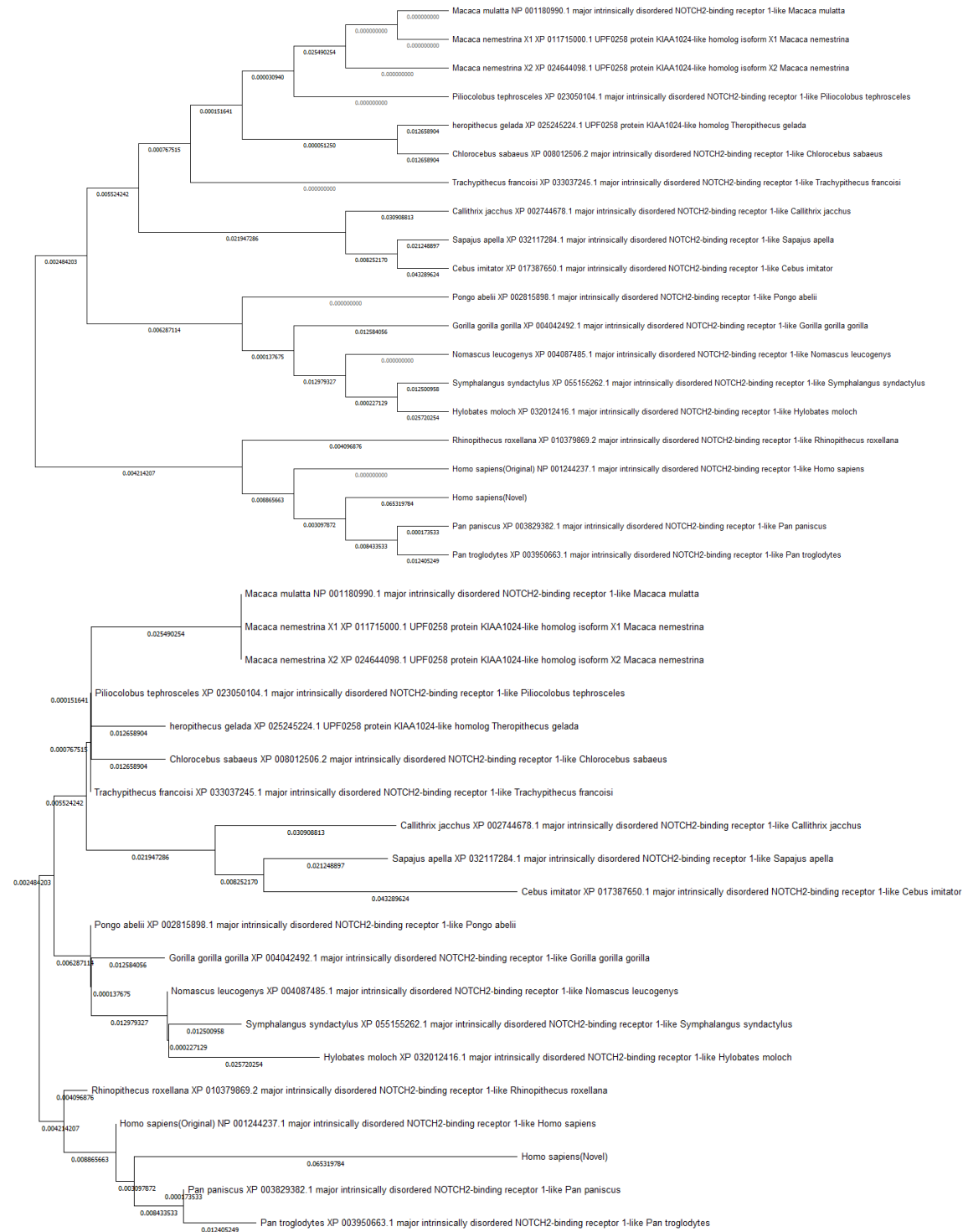
Q6. Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use "simple phylogeny" online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Sequences were imported to Mega11, aligned with MUSCLE, and neighbor-joining tree was generated. Due to collapsing of lines of subgroups at the top, both scale/non-scaled graphs are attached below.

Q7. Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and "Save as" FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.

```
library(bio3d)
library(pheatmap)

#Read aligned sequences
align_path <- "C:/Users/louis/Downloads/bimm143/R code/project/code"
fst_files <- list.files(path = align_path,
                        pattern = "*.fst",
                        full.name = TRUE)

basename(fst_files)

align <- read.fasta(fst_files, rm.dup = TRUE, to.upper = FALSE, to.dash=TRUE)


#Create heatmap
identity <- seqidentity(align, normalize=TRUE, similarity=FALSE, ncore=1, nseg.scale=1)
pheatmap(identity)
```
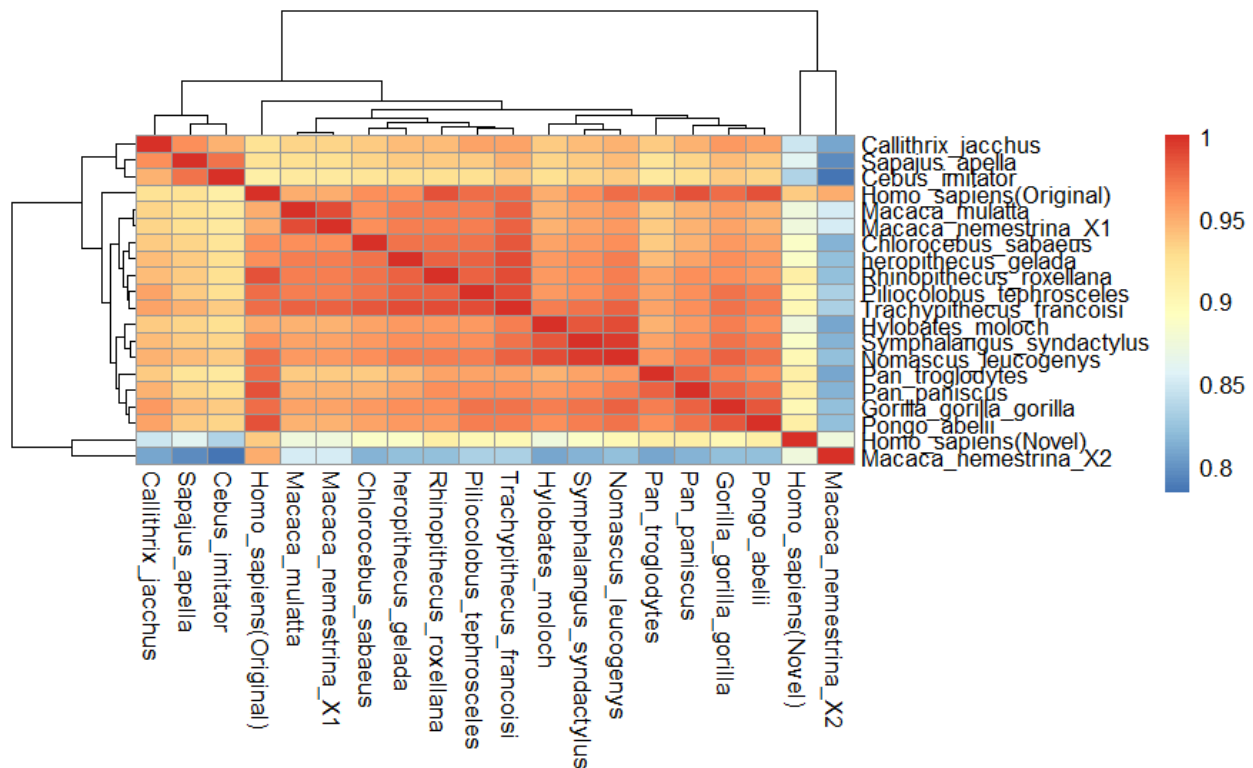
Q8. Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

```{r}
#Find consensus sequence
con <- consensus(align, cutoff = 0.9)
con$seq

#Run BLAST
blast.pdb(con$seq)

#Annotate BLAST result
ids <- c("3S2C_A", "4ATW_A", "3UG3_A")
pdb.annotate(ids, unique=TRUE)
```

| ID	Chain ID	Technique	Resolution	Source	Evalue	Identity (%)	Query cover (%)
3S2C	A	X-ray diffraction	3.0	Thermotoga petrophila RKU-1	9.4	46.667	23
4ATW	A	X-ray diffraction	3.0	Thermotoga maritima MSB8	9.3	46.667	23
3UG3	A	X-ray diffraction	1.8	Thermotoga maritima	9.3	46.667	23

Q9. Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black). Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

```
#Read pdb file for VMD
pdb.3S2C <- read.pdb("3s2c.pdb")

#Trim chain A
chainA <- trim.pdb(pdb.3S2C, chain = "A")
write.pdb(chainA, file = "chainA_3S2C")
```
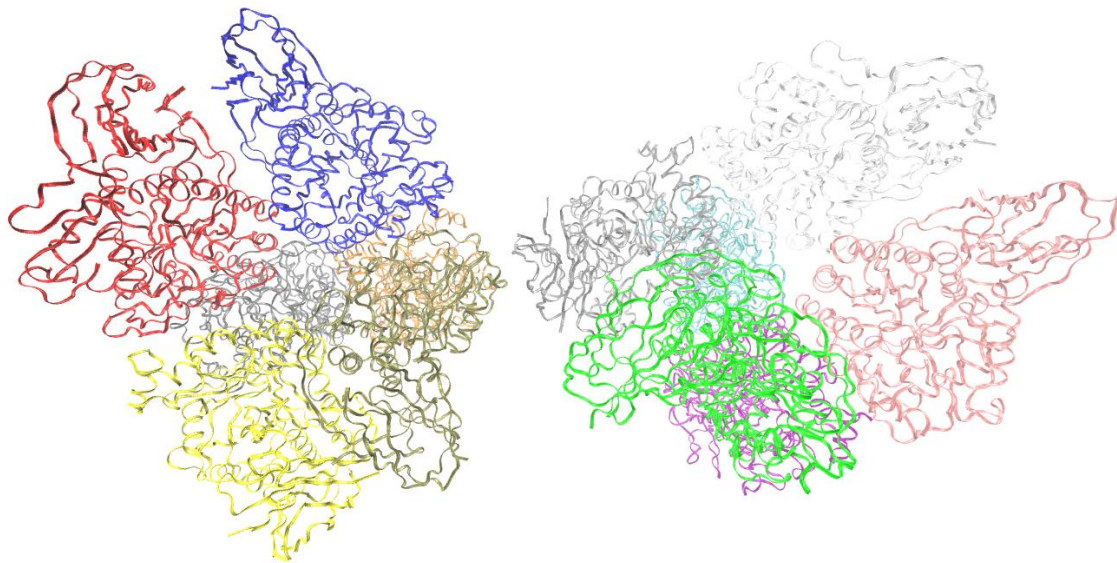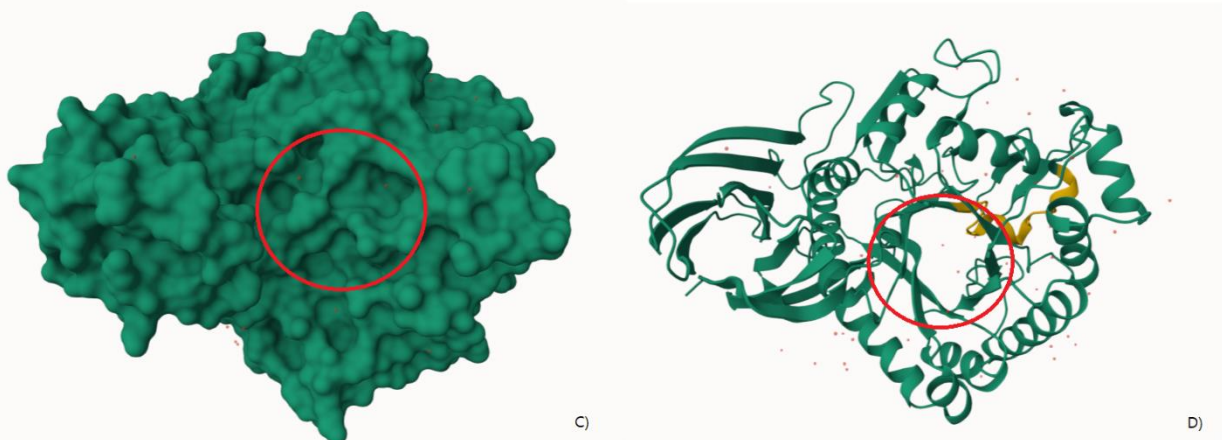

VMD view of 3S2C. Chain A, top match of blastp, is colored with blue.
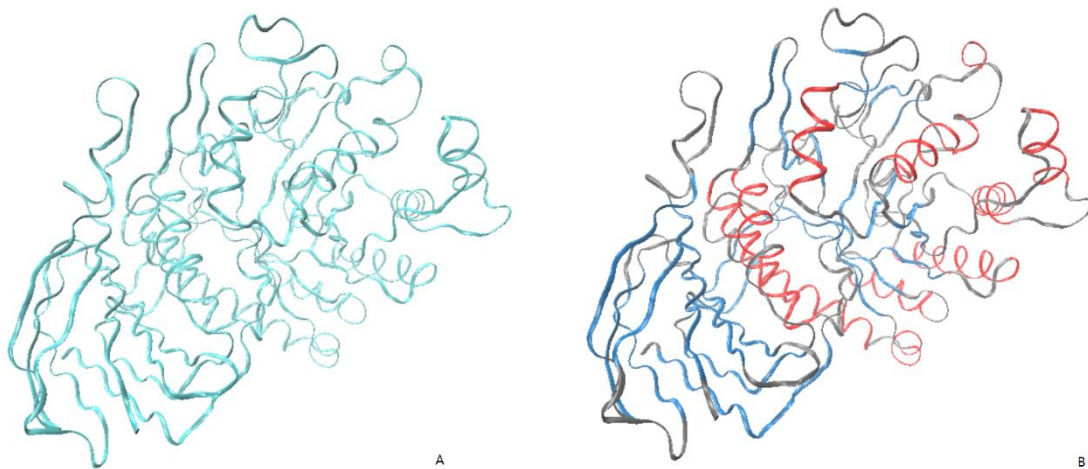
A: VMD view of 3S2C Chain A of 3S2C. B: Colored based on the secondary structure. Red – alpha helix, blue – beta sheet, grey – no secondary structure.
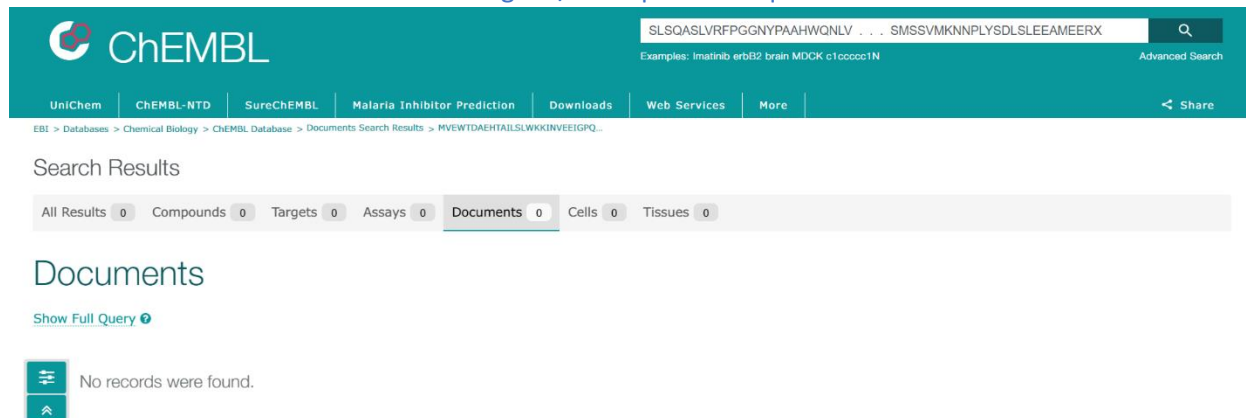
(Extra) C: Molecular surface of chain A in mol*. Red circle is where active site of 3S2C_A locates. D: Same view without molecular surface. Yellow beta sheet is where 3S2C_A matches with consensus sequence of MINAR2 proteins.

C-terminal Beta-sandwhich in multidomain protein (yellow in D) is essential in regulation of secretory process, anchoring substrates, and protein flexibility. Since MINAR2 is notch-oligomer binding protein, and 3S2C_A also binds to xylooligomer, functional similarity of oligomer binding suggests that novel protein is expected to have similar Beta-sandwhich structure.
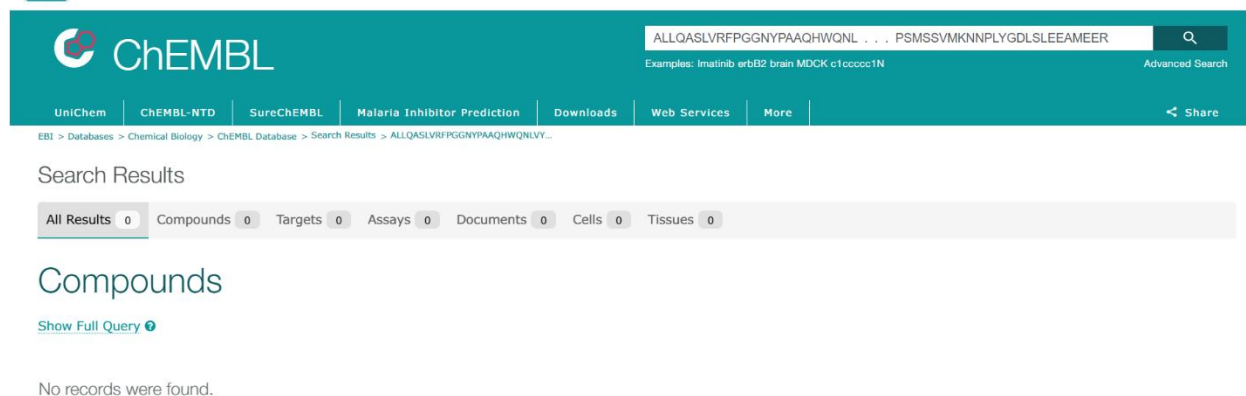
However, the low percent identity between novel-MINAR2 and 3S2C of 46.667% indicates that it is unlikely that 3S2C's structure will be similar to novel-MINAR2 based on sequence similarity. Also, considering that MINAR2 family is around 190aa, a two low coverage of 23% (15aa) out of short consensus sequence (65aa) further supports that novel-MINAR2 having similar structure to 3S2C is very unlikely.

Q.10 Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency ID Technique Resolution Source Evalue Identity 3BOM X-RAY DIFFRACTION 1.35 Oncorhynchus mykiss 6.59E-63 81.4 1SPG X-RAY DIFFRACTION 1.95 Leiostomus xanthurus 3.16E-58 75.9 3BCQ X-RAY DIFFRACTION 2.4 Brycon cephalus 5.11E-57 77.2 data reported that may be useful starting points for exploring potential inhibition of your novel protein?

There was no search result when both original/novel protein sequences were used.

Searching abbreviation for the protein resulted in unrelated information such as topoisomerase.

However, there were some results when full protein name (membrane integral NOTCH2-associated receptor 2) was searched: 1 assay (CHEMBL4040504) and 3 targets (CHEMBL1764937, CHEMBL3831223, CHEMBL2364701), even though no targets were exactly matching to MINAR2 family. All targets/assays showed no ligand efficiency data.

In scintillation proximity assay, displacing metabolic glutamate from human metabolic glutmate receptor 2 showed association with rate constant calculated from the assay (7-aryl-1,2,4-triazolo[4,3-a]pyridines

altering PAM activity), suggesting that high selectivity of glutamate receptor can be essential in optimal efficacy prediction. While this is not directly connected to MINAR2, it's meaningful in a sense that MINAR2 activity is associated with mTOR pathway, where mTOR is influenced by glutamine.

Citation:

Doornbos MLJ, Cid JM, Haubrich J, Nunes A, van de Sande JW, Vermond SC, Mulder-Krieger T, Trabanco AA, Ahnaou A, Drinkenburg WH, Lavreysen H, Heitman LH, IJzerman AP, Tresadern G. Discovery and Kinetic Profiling of 7-Aryl-1,2,4-triazolo[4,3-a]pyridines: Positive Allosteric Modulators of the Metabotropic Glutamate Receptor 2. J Med Chem. 2017 Aug 10;60(15):6704-6720. doi: 10.1021/acs.jmedchem.7b00669. Epub 2017 Aug 1. PMID: 28704052.