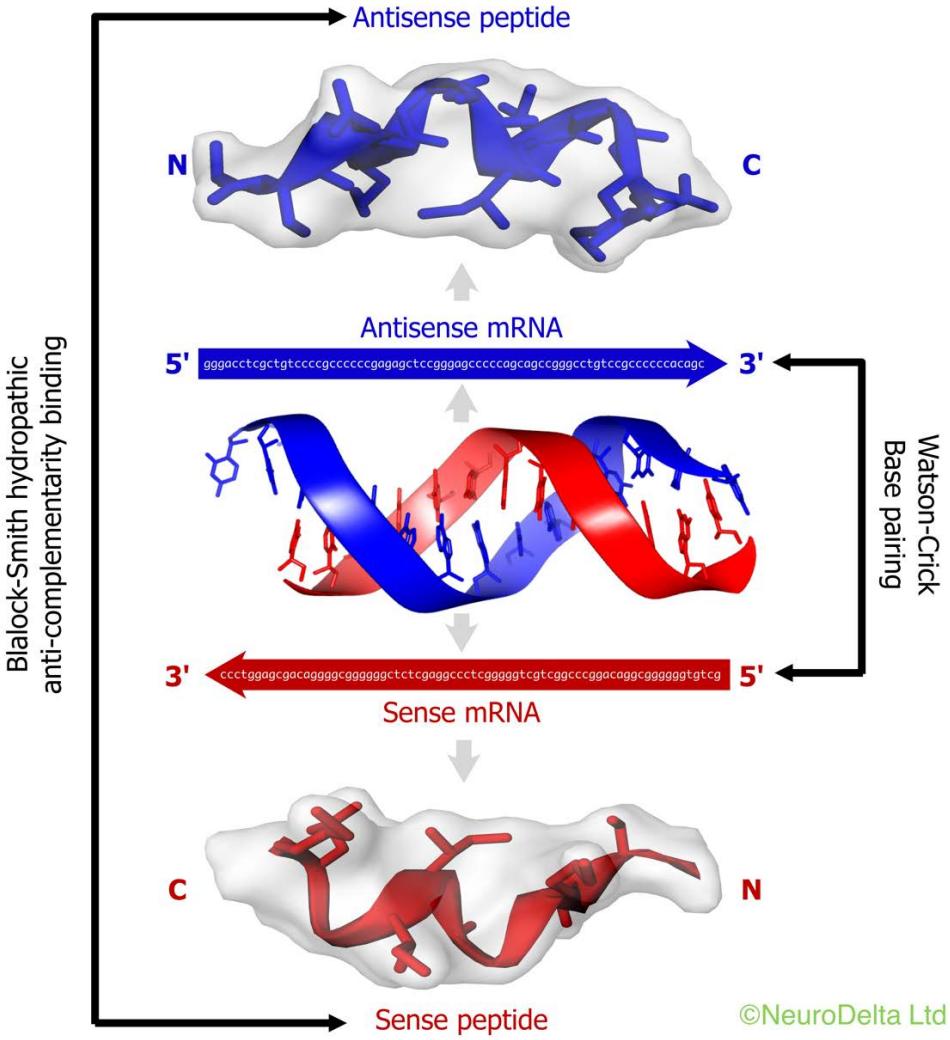
# Bioinformatics Protocols:
## Antisense Peptide with Molecular Modelling

# 1: Background, aims and method overview

**Background**:

Computer based prediction of protein-protein interactions is a valuable *in silico* tool (Capone *et al.,* 2008; McGuire & Holmes 2005; Siemion *et al.,* 2004; Štambuk *et al.,* 2018; Štambuk *et al.,* 2021) in a biological setting. Antisense peptide (or Complementary Peptides) sequences are derived from the complementary strand of DNA encoding a given protein (Bost & Blalock 1989a & 1989b), read in the same open reading frame (ORF). Due to the presence of exons and introns within the genomic DNA sequence the mRNA sequences are ideal for generating antisense peptides. They can also be derived directly from the amino acid sequence of a protein, via reverse translation to produce a complementary DNA sequence. However, due to the degeneracy of the genetic code, there is typically more than one antisense sequence for any one protein (Bost & Blalock 1989a & 1989b). The basis of antisense peptides based on the observations of Blalock & Smith (1984) is illustrated below:



The Molecular Recognition Theory is based on a series of observations that protein sequences derived from the Sense Strand of DNA bind to protein sequences derived from the corresponding Antisense Strand of the DNA (Biro 2007; Blalock & Bost 1988; Blalock & Smith 1984; Hardison & Blalock 2012; Heal *et al.,* 2002; Root-Bernstein & Holsworth 1998; Štambuk *et al.,* 2005; Štambuk *et al.,* 2021).

The complementary DNA strand for each individual amino acid can be read in either the forward 3'-5'or reverse 5'-3' direction, adding further degeneracy to the potential antisense peptide sequences (Bost & Blalock 1989a & 1989b; Milton 2006). The antisense peptides have been shown to bind with high affinity to the given target protein due to hydropathic interactions (Illingworth *et al.,* 2012; Pullen *et al.,* 2013), which is basically an interaction between water loving and water hating amino acids (Kyte & Doolittle 1982).

This protocol details the generation of antisense peptide sequences against a selected target (Bost & Blalock 1989b) and then using them to identify possible protein-protein interactions (Miller 2015). The method uses a python script to generate the antisense peptides against the selected target based on the methods of Milton (2006) and then BLAST searches to identify sequences with similarity to the antisense sequences that may interact with the target protein. This is followed by *in silico* modelling of protein-protein interactions between the target and identified protein from the BLAST search (Pierce *et al.,* 2011 & 2014).

Binding of proteins to their antisense proteins has been demonstrated in a number of studies and the antisense peptides have also been shown to have sequence similarity to receptor binding sites plus compounds that specifically bind the sense peptides (Blalock & Bost 1986; Bost *et al.,* 1985; Clarke & Blalock 1990; Fassina *et al.,* 1989; Milton *et al.,* 2001; Mulchahey *et al.,* 1986; Štambuk *et al.,* 2019). The antisense peptides themselves have been used as binding peptides to modify the actions of the target protein (Bost *et al.,* 1985; Štambuk *et al.,* 2021). Antibodies raised against antisense peptide sequences have also been used to identify binding proteins *in vitro* or *in vivo*, for example LHRH as shown by Mulchahey *et al.,* (1986).

From a Bioinformatic point of view the DNA sequences of the Sense Strands that encode proteins are contained in many databases including the NCBI Nucleotide Database (https://www.ncbi.nlm.nih.gov/nucleotide/). As such it is possible to download the DNA sequence and use this as a source material. The amino acid sequences of the proteins from many species are likewise contained in many databases such as NCBI Protein Database (https://www.ncbi.nlm.nih.gov/protein/). If the sequences for proteins encoded by the Antisense strand of DNA for a given target protein are derived, they can be used to search protein databases for similar proteins using sequence comparison tools (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

The nature of the identified interactions from this process is two-dimensional and does not fully take into account the three-dimensional structure (3D) of each protein. The final stage of this protocol is to use 3D modelling to predict the structure of complexes between the target protein and binding protein identified due to sequence similarity with the antisense peptide derived from the target protein mRNA. The original ZDock method (Chen & Weng 2002) used for this has been updated and is well described by

Pierce *et al.,* (2011 & 2014). The method uses an algorithm to predict the protein-protein interactions with input of data from the identified interactions. The combination of the antisense methods with 3D modelling strengthens the *in silico* observations. Key confirmation of the validity of identified interactions is to determine that the *in silico* findings can be replicated *in vitro* and/or *in vivo*, for example catalase binding to amyloid-ß (Milton *et al.,* 2001) and the role of this interaction in preventing the toxicity of amyloid-ß combined with demonstration in samples from an Alzheimer's patient (Chilumuri *et al.,* 2013a & b).

**Aims**:

(i) Generation of Antisense peptide sequences for a chosen target protein.

(ii) Identification of potential protein binding partners for the chosen target protein.

(iii) Identification of the regions of each interacting protein pair that are directly involved in the binding.

(iv) 3D modelling of identified interactions

**Method Overview**:

The method uses either a PC or Mac (both Intel and Apple Silicon powered) based computers and can be run on Windows 10/11 or Mac OS. The protocols have been written using Microsoft Office 365 software with Microsoft Word and Microsoft Excel the main programs employed. The Python program for antisense peptide generation is compatible with Python 3 and has been tested on versions 3.7 upwards. Antisense peptide generation can also be carried out using Microsoft Word if online Python or Python installation is not possible.

(i) Generation of antisense peptides (Milton 2006) that are model binding proteins for the chosen target.

(ii) Searching of protein databases using a BLAST search for proteins that are similar to the antisense sequences and therefore may also be binding proteins for the chosen target.

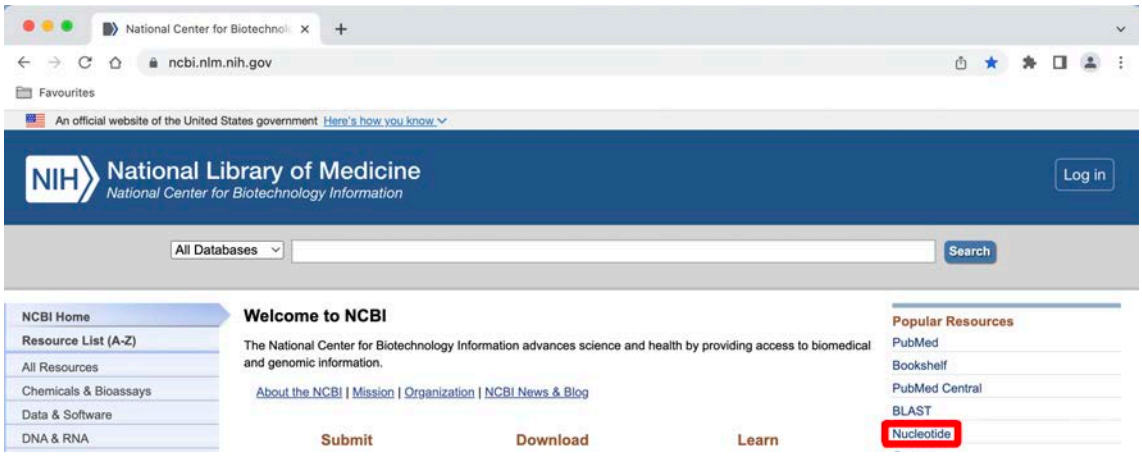
(iii) Extraction of data from the BLAST search and exclusion of results that are incompatible with the protein binding, for example sequences with Gaps that are identified in by the BLAST searching techniques, but which are would not indicate a potential binding interaction (Pullen *et al.,* 2013).

(iv) Scoring of the binding interactions based on the Molecular Recognition Theory (Hardison & Blalock 2012) to identify the most relevant protein-protein interactions from the BLAST search.

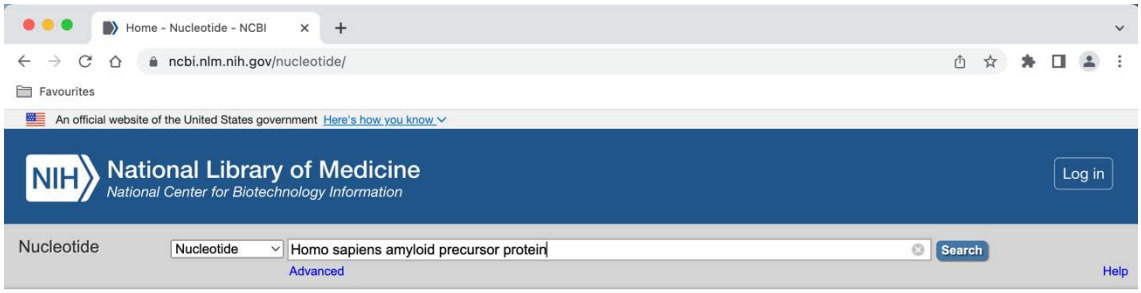
(v) 3D molecular modelling of the potential protein-protein interactions (Pierce *et al.,* 2011 & 2014) using the potentially interacting residues identified in BLAST searches as a basis for the contacting residues in the model.

# 2: Target protein mRNA search

(2a) Stage one is to find the mRNA coding sequence for the protein of interest. Best to use a normal browser such as Microsoft Edge, Safari, Google Chrome or Firefox for these searches. The mRNA sequences can be found on the https://www.ncbi.nlm.nih.gov databases – click on the nucleotide link (circled in red).

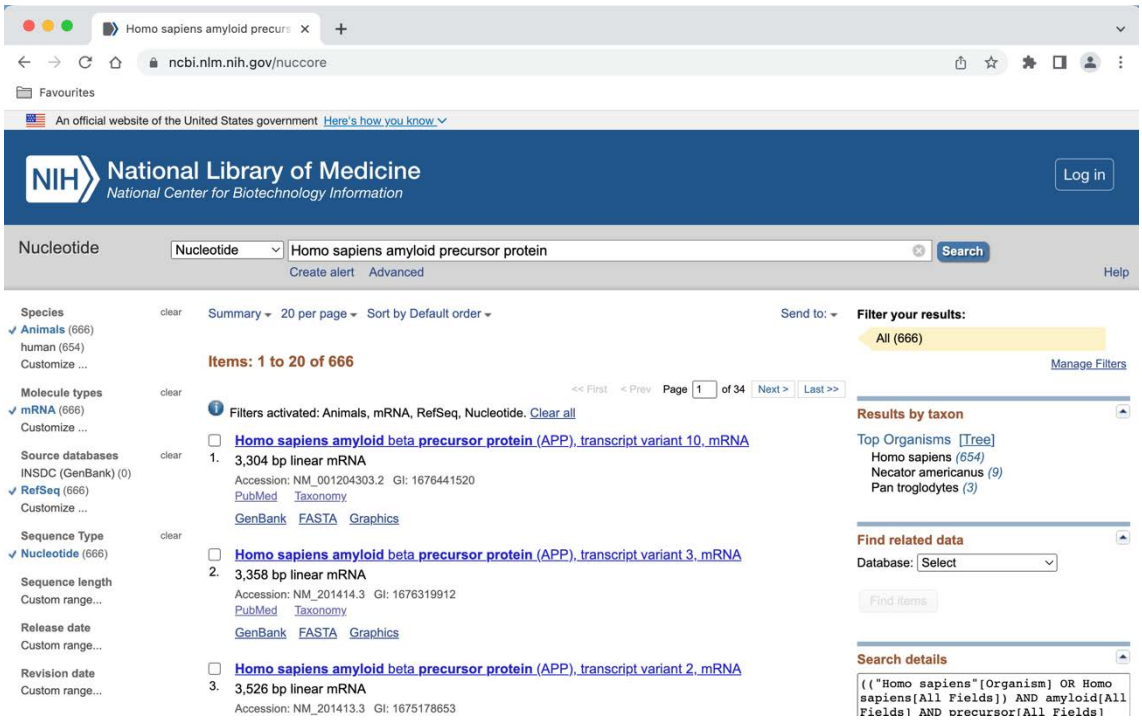



Which will go to where the name of protein of interest can be entered (as an example the Homo sapiens amyloid precursor protein has been used:

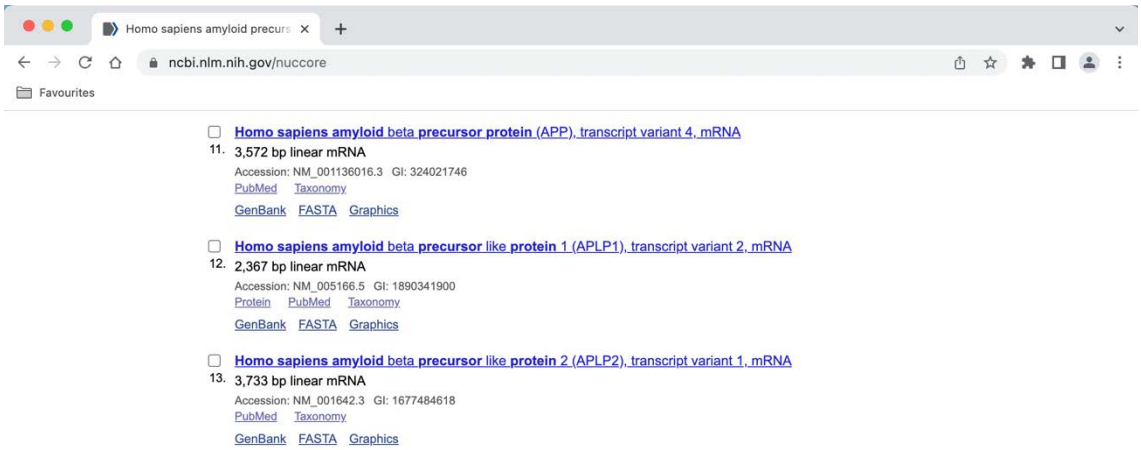



This will then go to a list of nucleotide sequences that match the search term. By selecting Animals, mRNA and Nucleotide on the Left-hand side (marked in blue with a tick after selection – the sequence list can be limited to those of most use. Selecting RefSeq limits the search further to those sequences that have been reviewed by the NCBI and is therefore recommended. The main features of the RefSeq collection include:

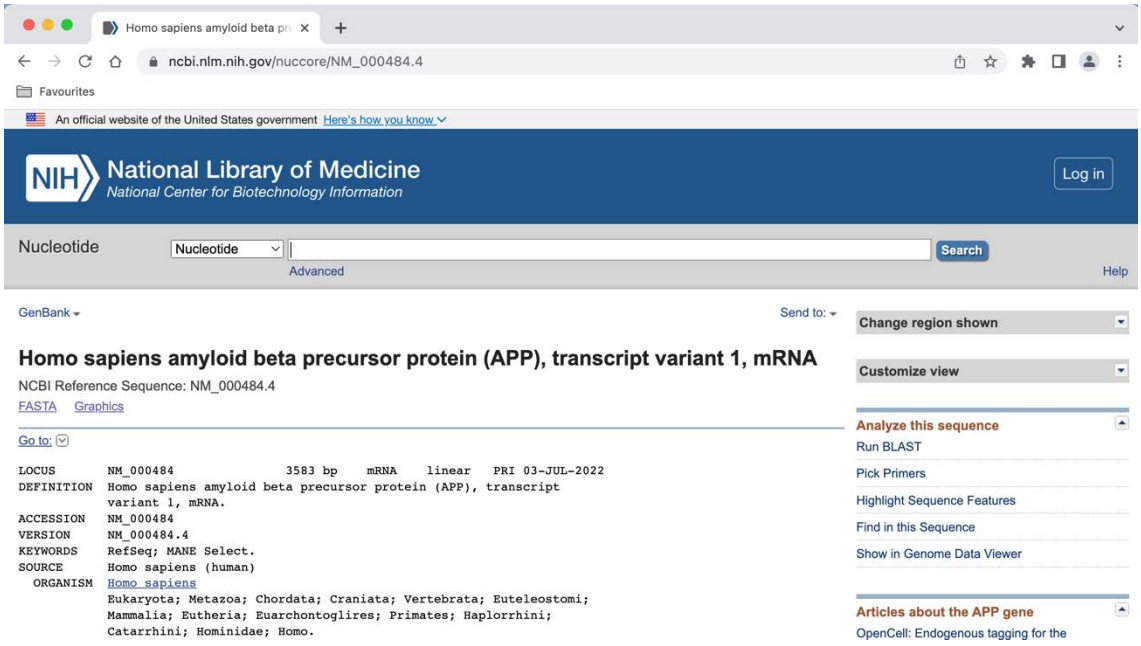
- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- distinct accession series (all accessions include an underscore '_' character)
- ongoing curation by NCBI staff and collaborators, with reviewed records indicated



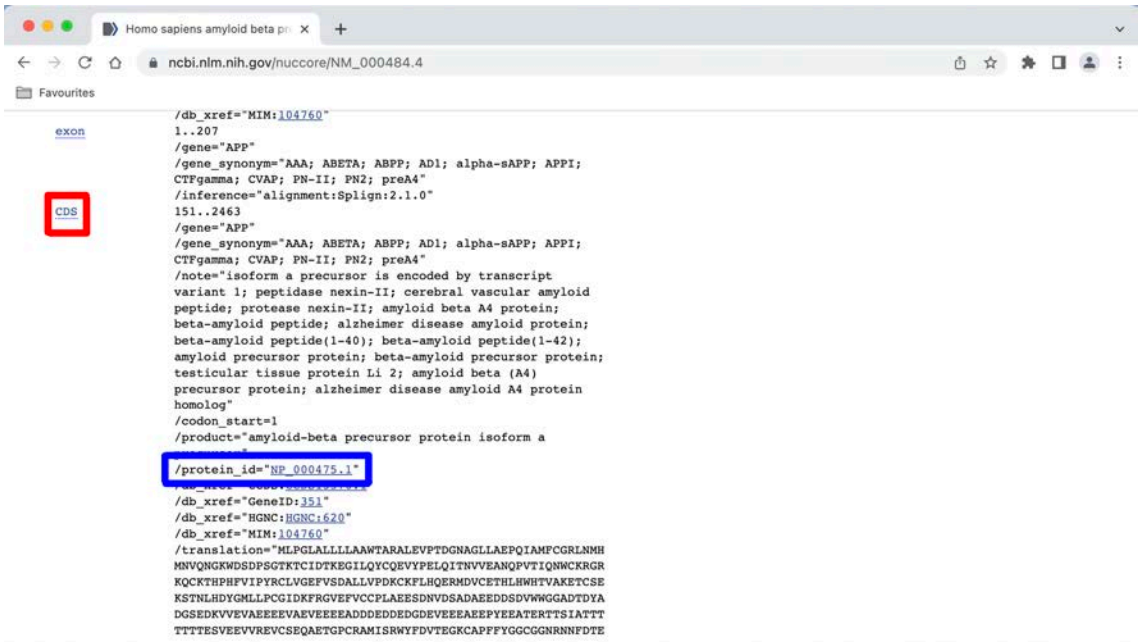

The choice of sequence is dependent on several features. If the literature suggests a particular transcript is of interest, then that is the recommended sequence to choose. It is also important to check the names carefully, in the example below the search term "Homo sapiens amyloid precursor protein" will also identify related proteins such as the "amyloid precursor like proteins":



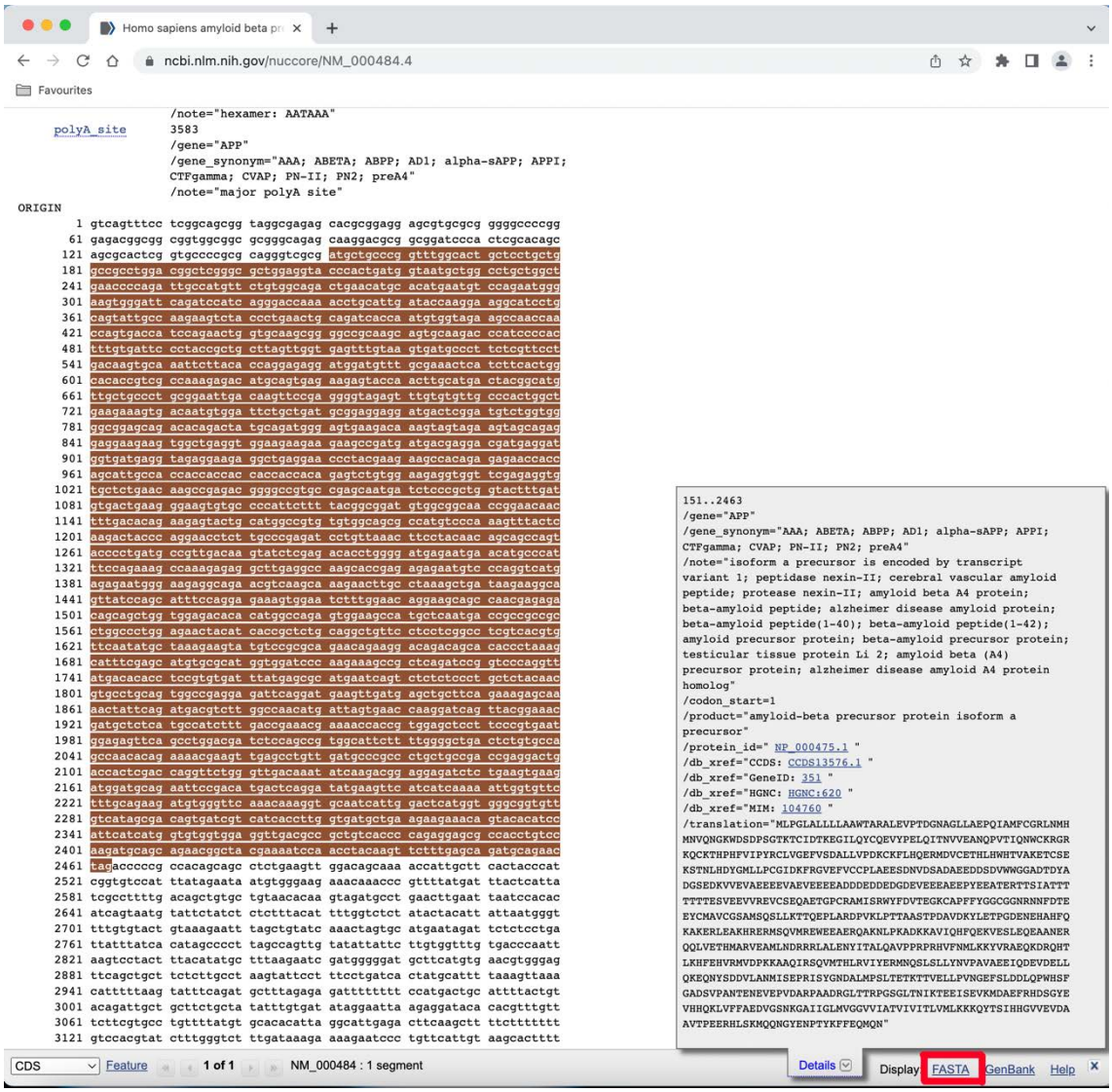

The transcript of interest should be selected, generally select the longest transcript and run that through the antisense peptide generation first. Clicking on the link will go to:



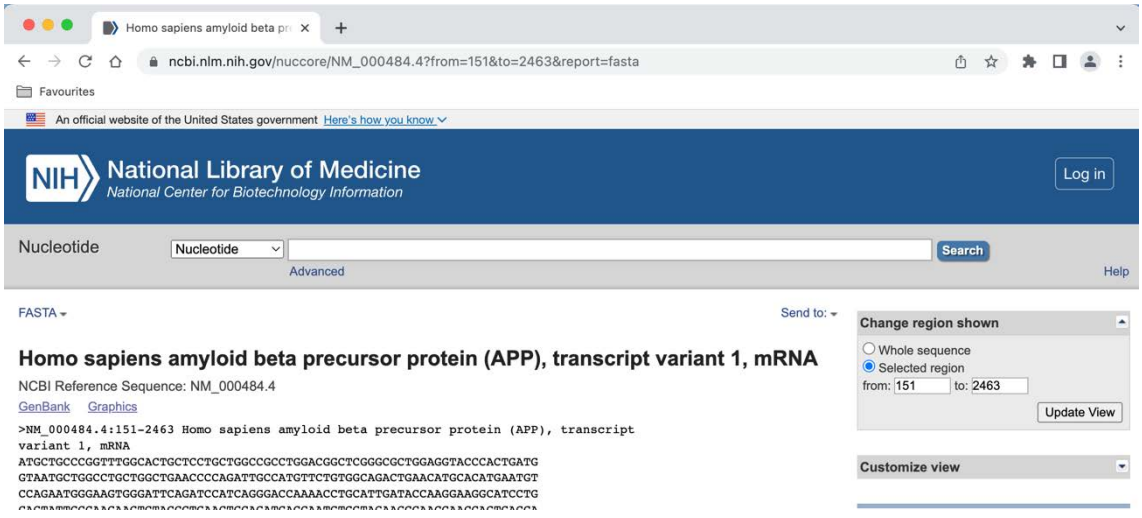

The next step is to scroll down this page until the CDS link (circled in red) is reached:




(2b) Clicking the protein id link (circled in blue, = NP_000475.1 in this example) will go to the coded protein and this number (normally starting NP_) can also be used in the BLAST check detailed in Section 4a (pages 15-16). Click on the CDS link (circled in red) will go to the following:

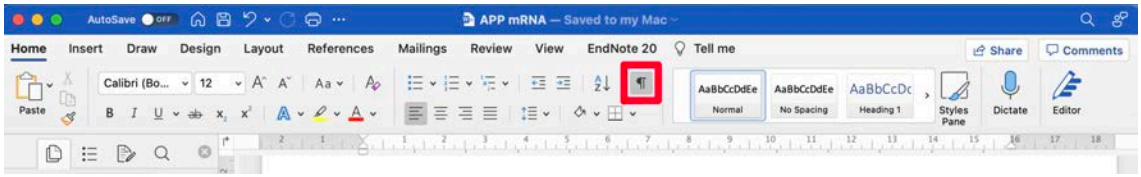



The part in Brown is the part required to generate antisense peptides. The FASTA tab (circled in red) will go to:



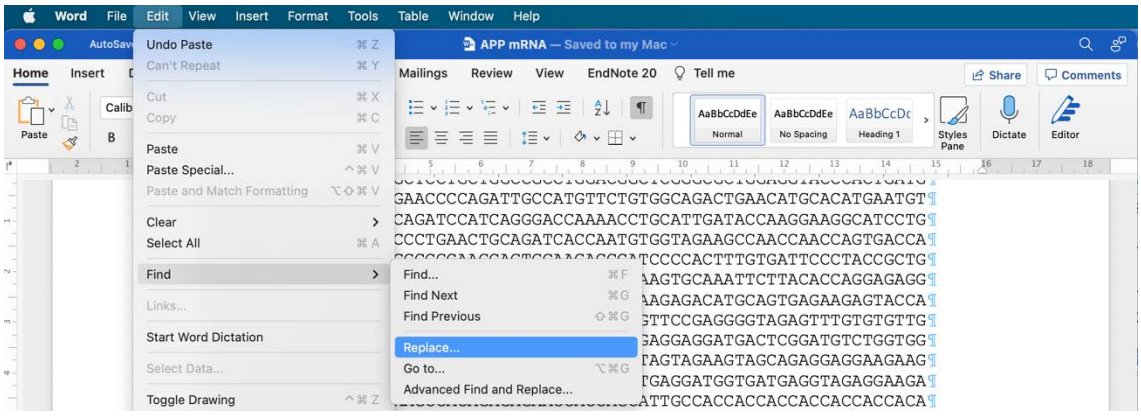
(2c) The text starting at ATG and finishing at either TAA, TAG (in this example - https://www.ncbi.nlm.nih.gov/nuccore/NM_000484.4?from=151&to=2463&report=fasta) or TGA should be copied to a word file and save as mRNA. This is the sequence that will be used by the AntisensePeptide.py program to generate the antisense sequences after removal of the returns.

The returns at the end of each line are removed as follows. First in word select the view symbols tab (circled in red below):

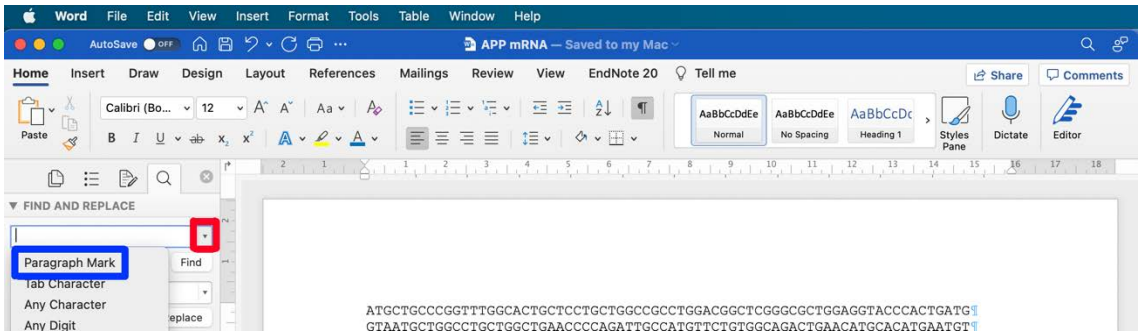



This will show the returns as "¶", which need to be deleted:

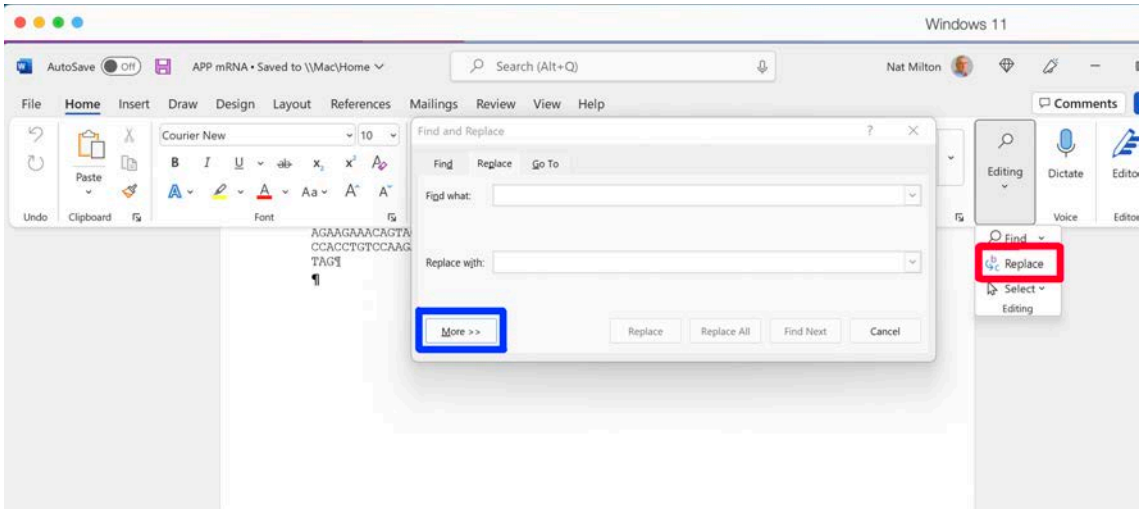

In the file go to the EDIT tab, then Find and select Replace:



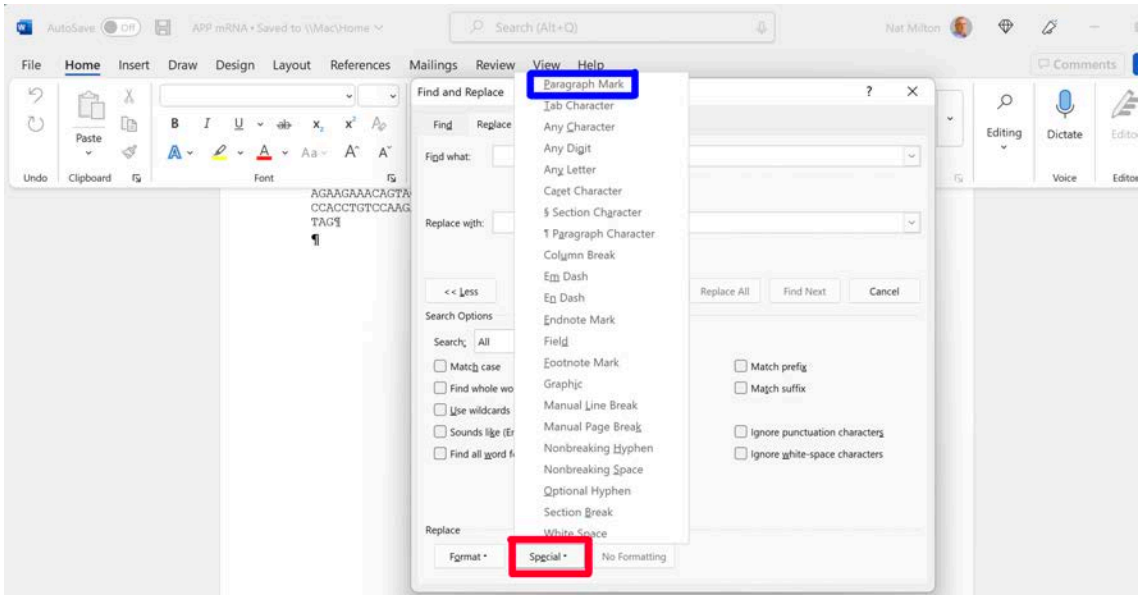

For Mac this shows as follows, when the arrow in the Find box is clicked (circled in red), then select Paragraph Mark (circled in blue) and leave the Replace blank:




On a PC in Word running in Windows the trick is to select replace (crtl H – also circled in red) and then the More button (circled in blue):

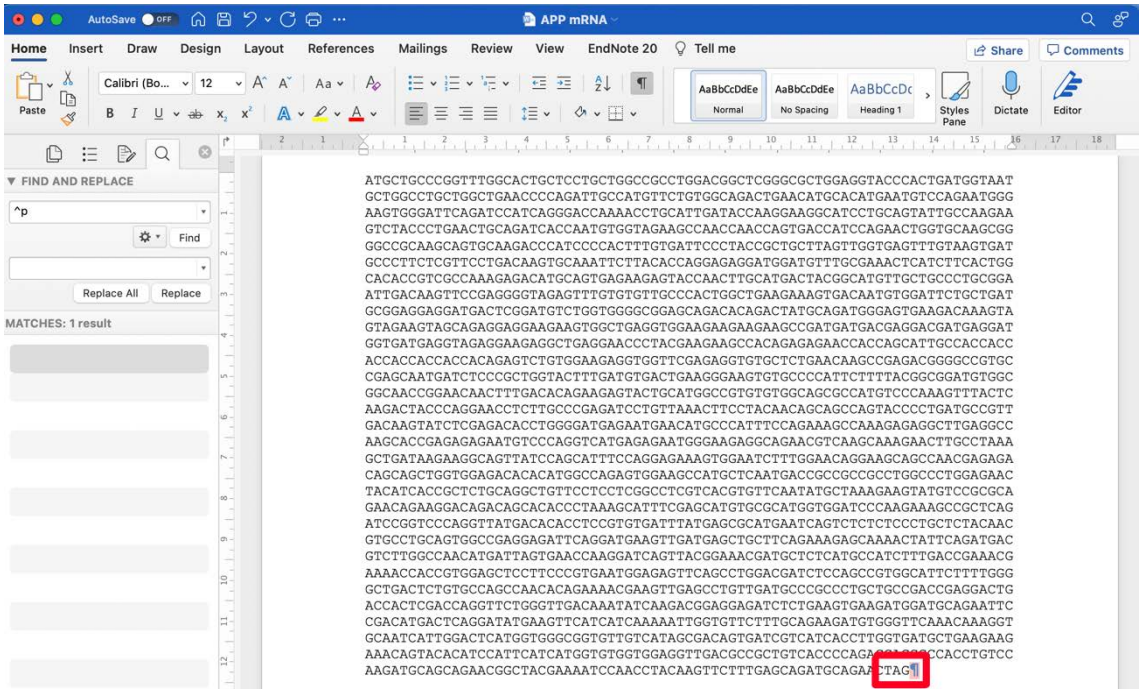



Then with the More window open click on the Special tab (circled in red) and then select the Paragraph Mark will then be displayed at the top (circled in blue):



This will then show as ^P (circled in red) in the find box and blank in the replace box and the Replace All (circled in blue) should then be clicked to remove all the "¶" symbols at the end of each line:


A single "¶" symbol at the end of the sequence after the TAA, TAG or TGA (in effect the whole sequence as a single word) will remain and the TAA, TAG or TGA (circled in red) triplet can be deleted as these stop codons are not needed:

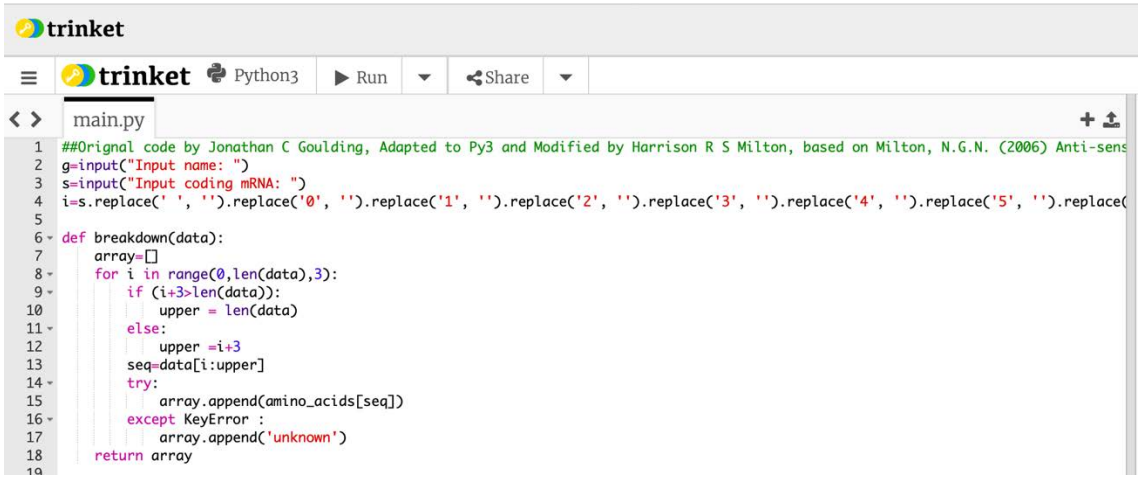



(2d) Then save file and use this version of the mRNA sequence for Antisense peptide generation as described in either Section 3 below using the online Python 3 compiler (pages 12-14 below), using the downloaded Python 3 program (see Section 14 pages 80-83 and Section 15 pages 84-87) or manually as detailed in Section 16 (pages 88-92).

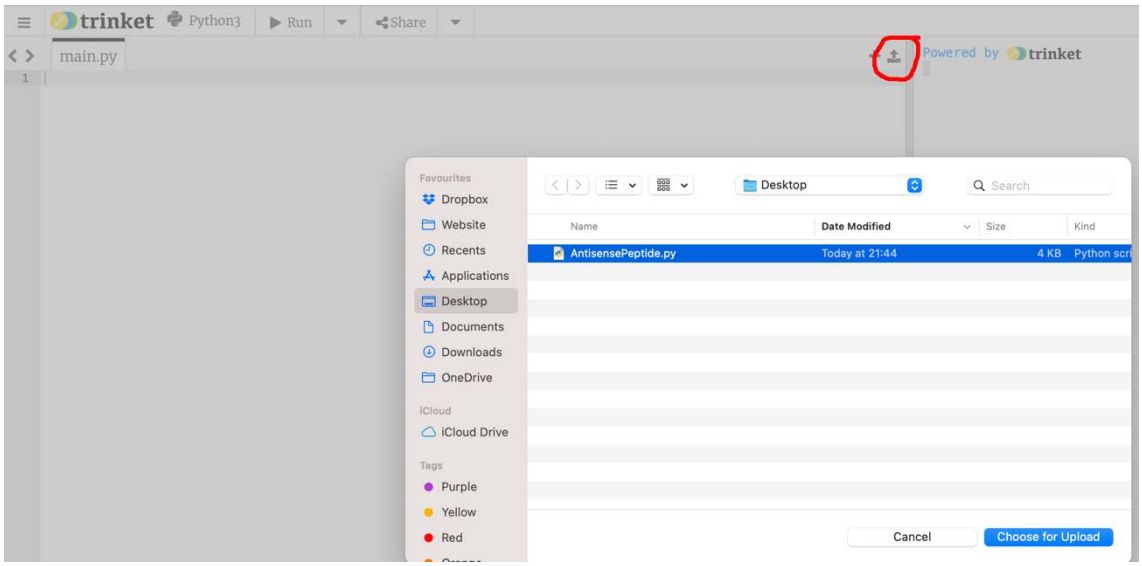
# 3: Antisense peptide generation using online Python

(3a) An online Python compiler (https://trinket.io/python3) can be used to generate antisense peptide sequences using the Antisense-Peptide.py file. This can be run in most browsers on a Mac, PC, Chromebook, iPad etc and has been tested using both Safari and Google Chrome.

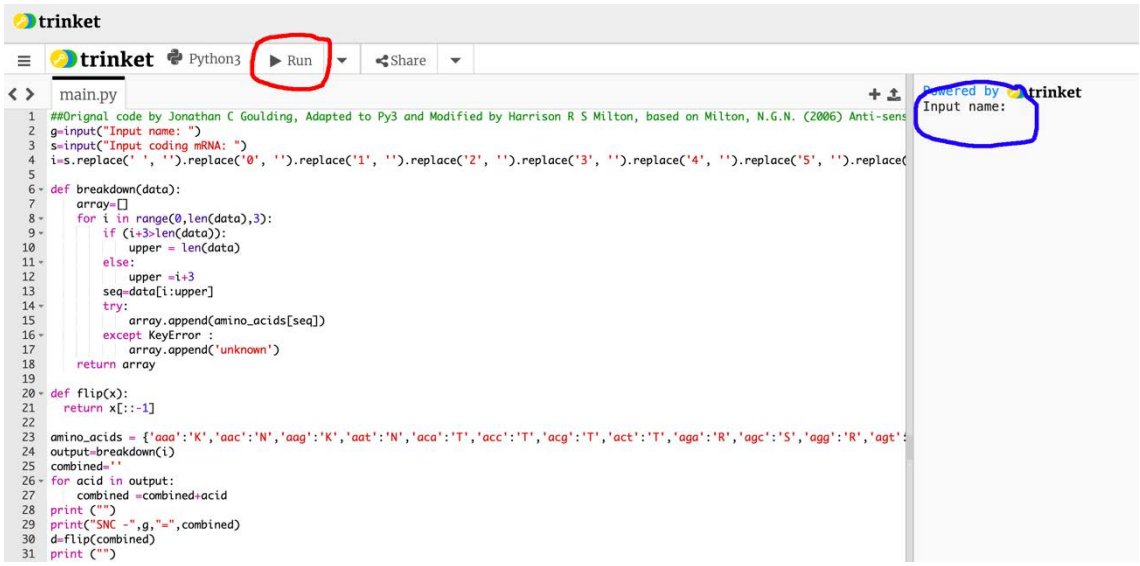
(3b) The text of the Antisense-Peptide.py (available from as a either Python script https://www.neurodelta.uk/resources/BioinformaticsProtocolScript.py or as a Word file from https://www.neurodelta.uk/Protocols/) can be copied and pasted into the compiler:



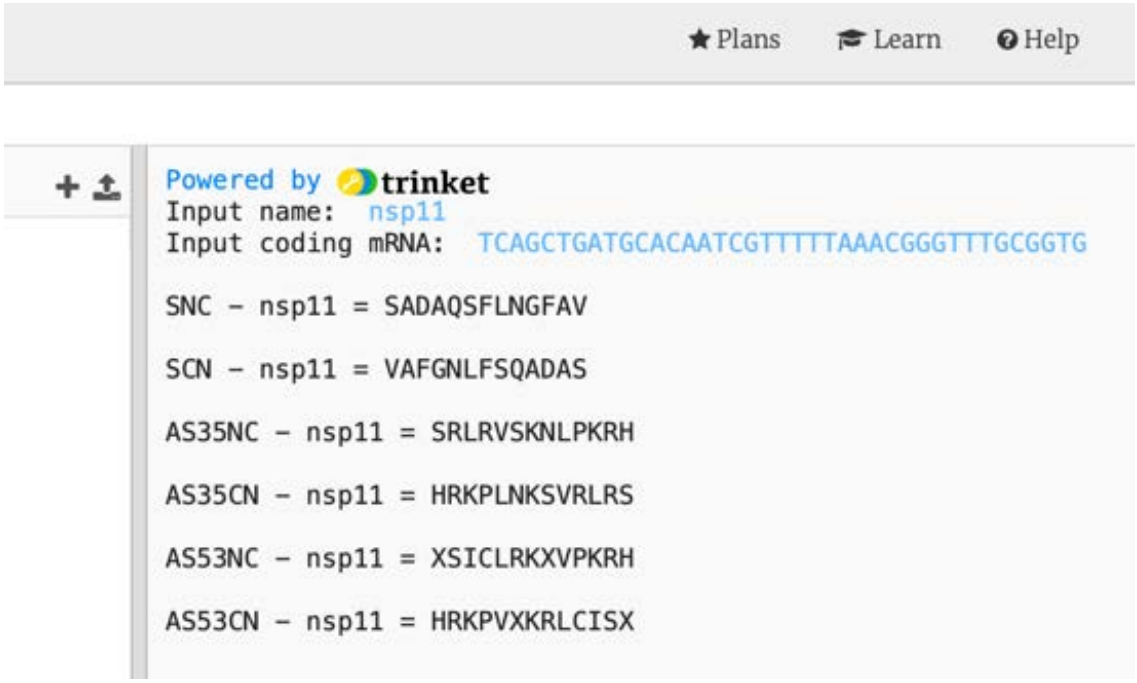

(3c) The Antisense-Peptide.py Python Script can also be downloaded and saved to an appropriate location on a computer hard drive and then uploaded using the compiler upload function (circled in red) followed by selection of the Antisense-Peptide.py file:



(3d) The compiler can then be run using the run function (circled in red) which will bring up the "Input Name:" command (circled in blue), the name or abbreviation of the target protein without spaces should be typed here followed by a return:




This will bring up the "Input coding mRNA:" command, the mRNA sequence for the target protein prepared in section 2d (page 10 above), in this example the sequence TCAGCTGATGCACAATCGTTTTTAAACGGGTTTGCGGTG is used:




In a very rare number of cases the a, t, c or g residues in the mRNA sequence could be replaced by an "n". This will cause an UNKNOWN to show in the peptide sequences which should be replaced by an X.

Where there is an * at the start (SCN, AS35CN and AS53CN) or end (SNC, AS35NC, AS53NC) of a sequence this is where the STOP codon was in the mRNA

and can be deleted form the sequences used to run BLAST searches. If there is an * or an UNKNOWN in the middle of a sequence this indicates a problem with the mRNA used as these should only be at the end of coding sequences. Suggests a need to repeat section 2a-2d (see pages 5-10 above) to get the correct CDS mRNA component, particularly check section 2c and 2d (see pages 8-10 above) to create an mRNA sequence that is a single word has been completed properly.

(3e) Copy the text from Input name down to the end of the AS53CN sequence and paste into a word document:

Input name: nsp11
Input coding mRNA: TCAGCTGATGCACAATCGTTTTTAAACGGGTTTGCGGTG

SNC - nsp11 = SADAQSFLNGFAV

SCN - nsp11 = VAFGNLFSQADAS

AS35NC - nsp11 = SRLRVSKNLPKRH

AS35CN - nsp11 = HRKPLNKSVRLRS

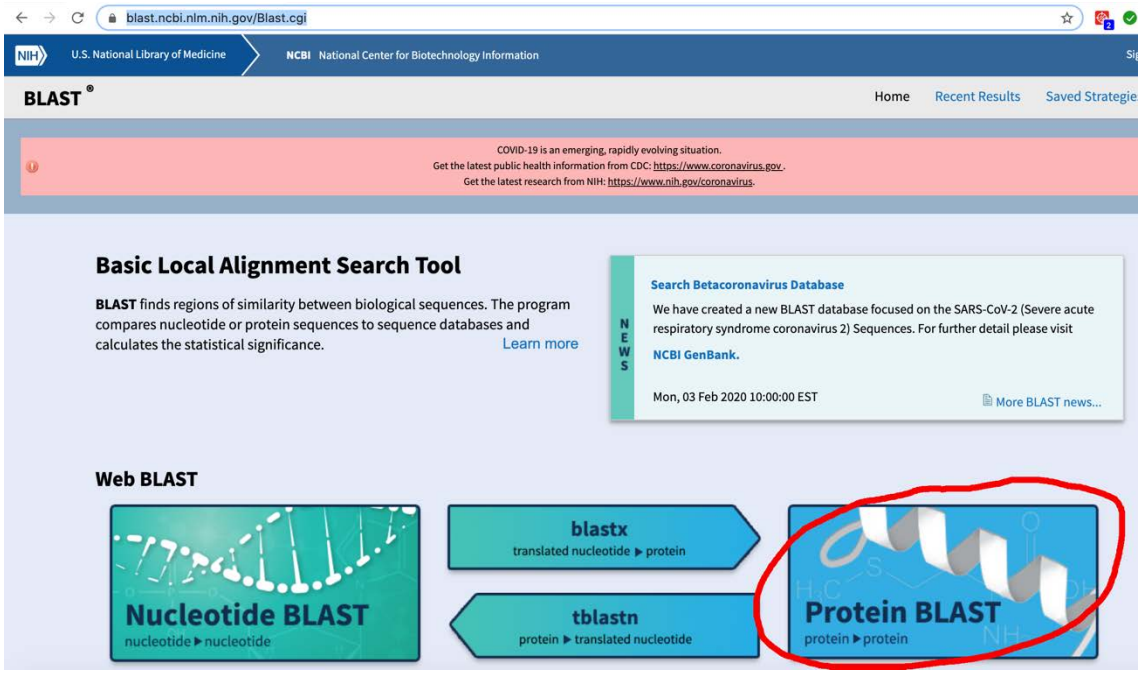
AS53NC - nsp11 = XSICLRKXVPKRH

AS53CN - nsp11 = HRKPVXKRLCISX

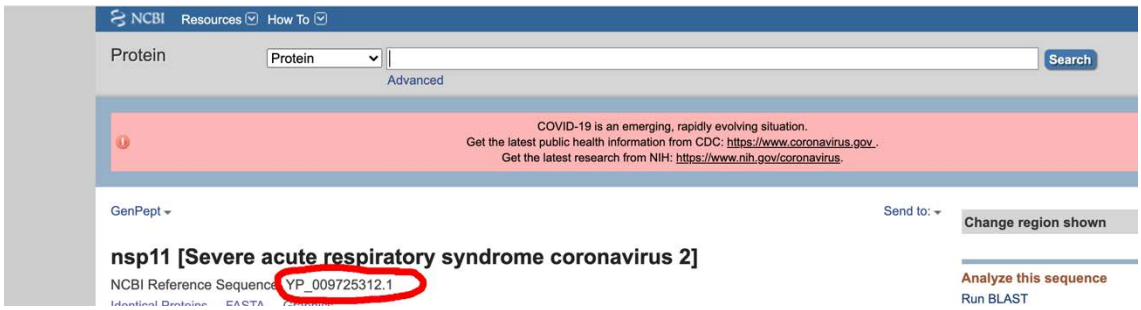
Save the Python outputs word file with a suitable name. These are the sequences that will be used for BLAST searches in section 4 (see pages 14-16 below) and section 5 (pages 17-19 below):

# 4: Confirmation of correct target mRNA

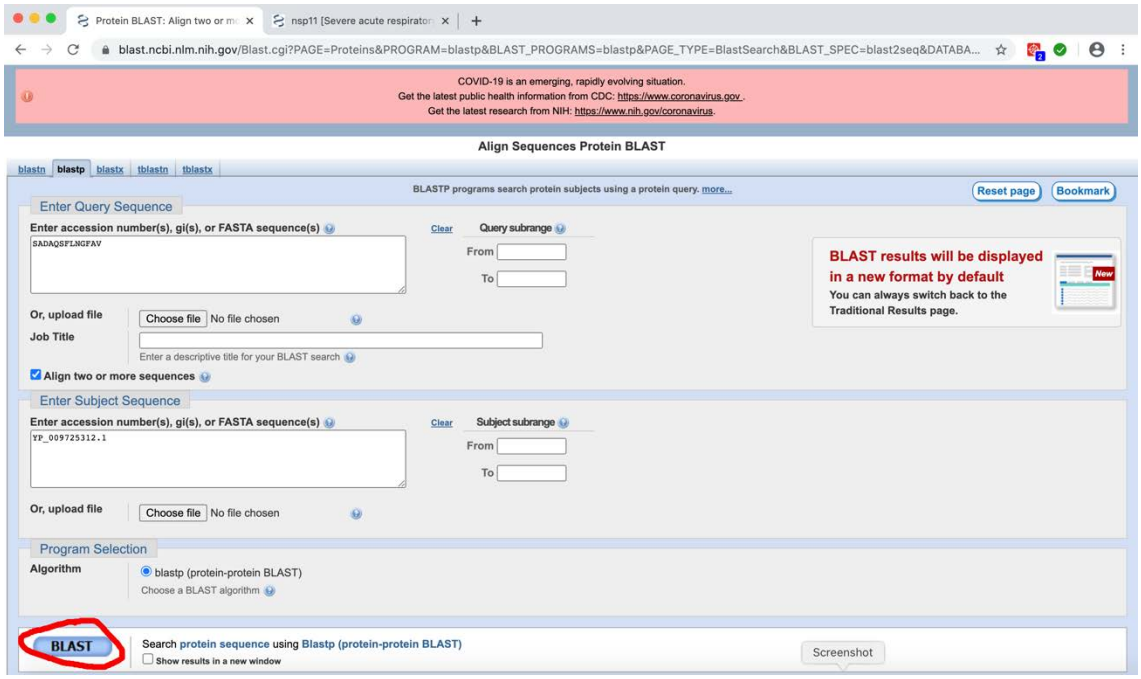
(4a) The SNC sequence from the running AntisensePeptide.py on Python is the sense sequence of the protein obtained from the mRNA sequence, to check that the correct mRNA sequence has been used in AntisensePeptide.py a BLAST search using the https://blast.ncbi.nlm.nih.gov/Blast.cgi website can be run in the browser allowing a search for sequence identities with the SNC sequence. Select Protein Blast (circled in red):



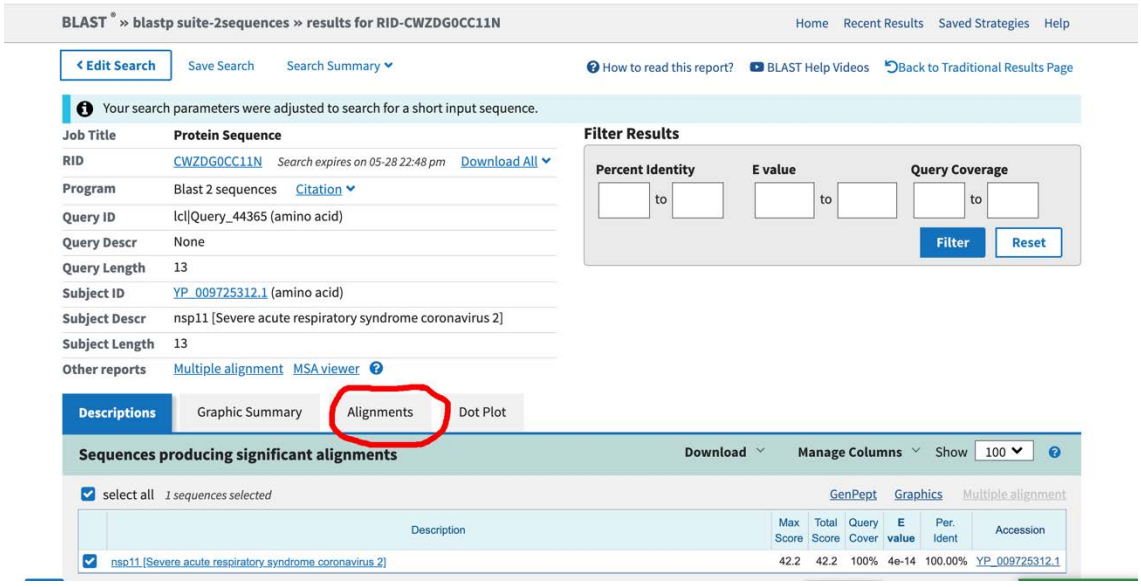

From within the Protein BLAST paste the SNC sequence into the Enter Query Sequence. There are a number of options to select from. If the protein id for the sequence for the mRNA used to generate the antisense peptides is available (see section 2a, page 6) the Align two or more sequences checkbox can be selected and the protein id number pasted into the second box. For the purposes of the example used in the AntisensePeptide.py program above the sequence needed to BLAST is SNC – nsp11 = SADAQSFLNGFAV. The protein sequence id for nsp11 is YP_009725312.1 (circled in red) from the search of the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein):



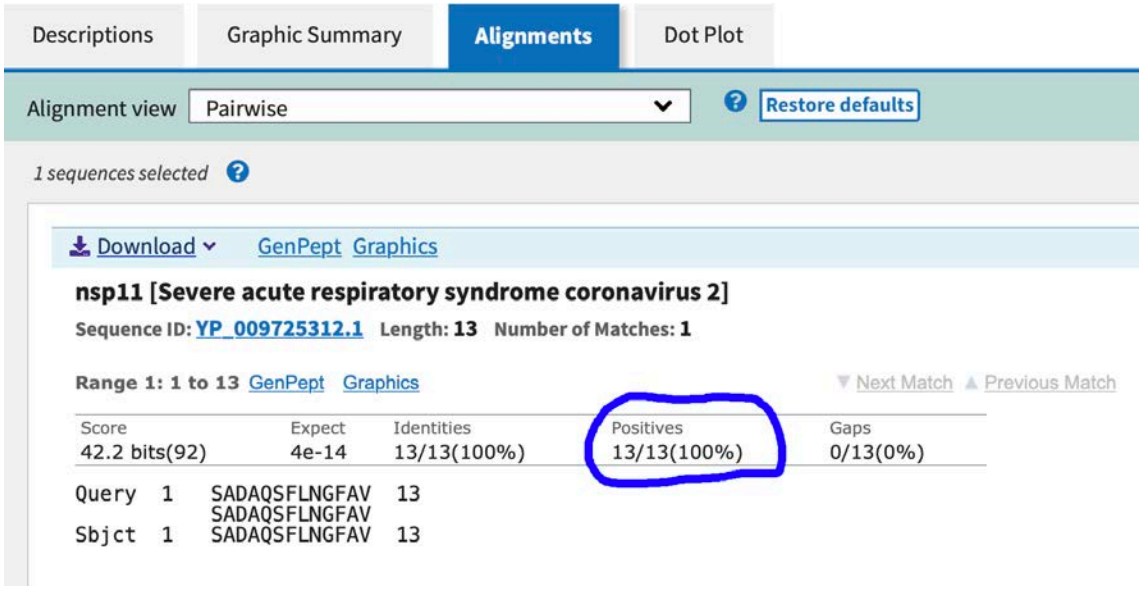
In the blastp search box enter the sequence, make sure the Align two or more sequences checkbox is ticked, paste the SNC sequence in the first box and enter the Protein id in the second box:




Clicking on the BLAST button (circled in red) will bring up the following:



(4b) Selecting the Alignments tab (circled in red) shows the comparison of the amino acids, in this case confirming the SNC and the original protein are identical and therefore that the correct mRNA was used in the Antisensepeptide.py program the key number is the Positives, which should be 100% (circled in blue):



# 5: Antisense peptide BLAST searches

(5a) The outputs form the AntisensePeptide.py program include the sense protein in normal N-terminus to C-terminus orientation (SNC) and also the sense protein in the reverse C-terminus to N-terminus orientation. The protein databases are always N-terminus to C-terminus orientation; however, proteins may interact with the binding site having one protein in the N-terminus to C-terminus orientation and the other in the C-terminus to N-terminus orientation. Hence the need to search the C-terminus to N-terminus orientation antisense peptides.

The antisense sequences generated are:

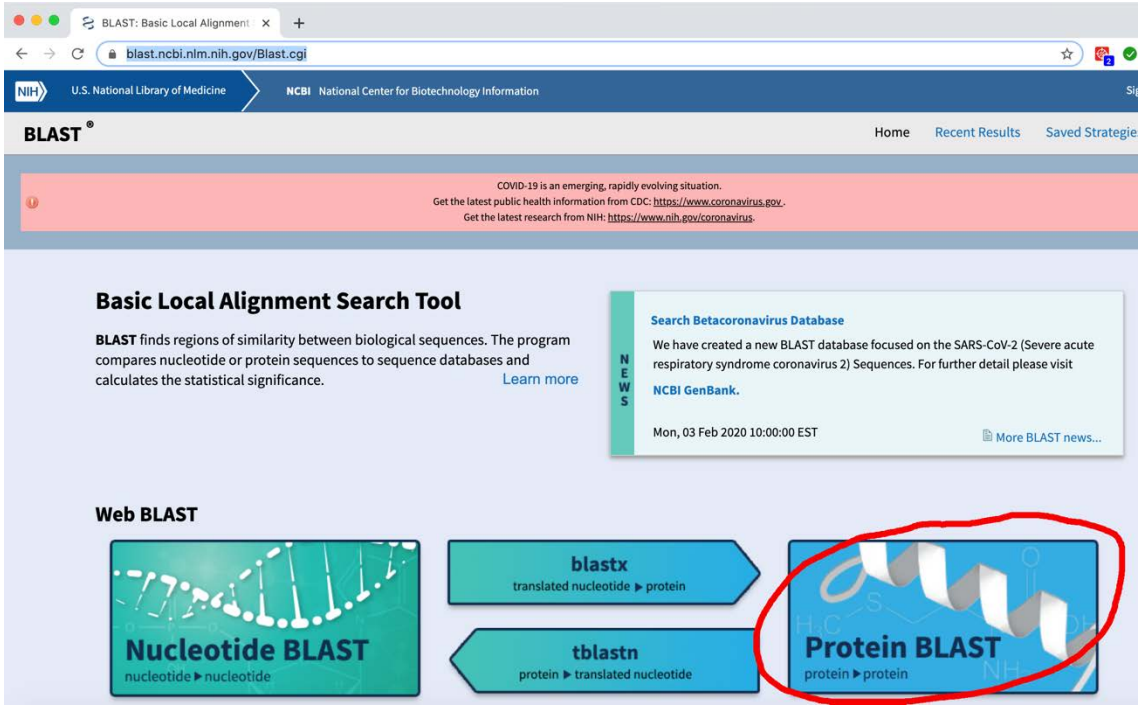
AS35NC = Antisense peptide (3'-5' mRNA reading) N-terminus to C-terminus orientation

AS35CN = Antisense peptide (3'-5' mRNA reading) C-terminus to N-terminus orientation

AS53NC = Antisense peptide (5'-3' mRNA reading) N-terminus to C-terminus orientation

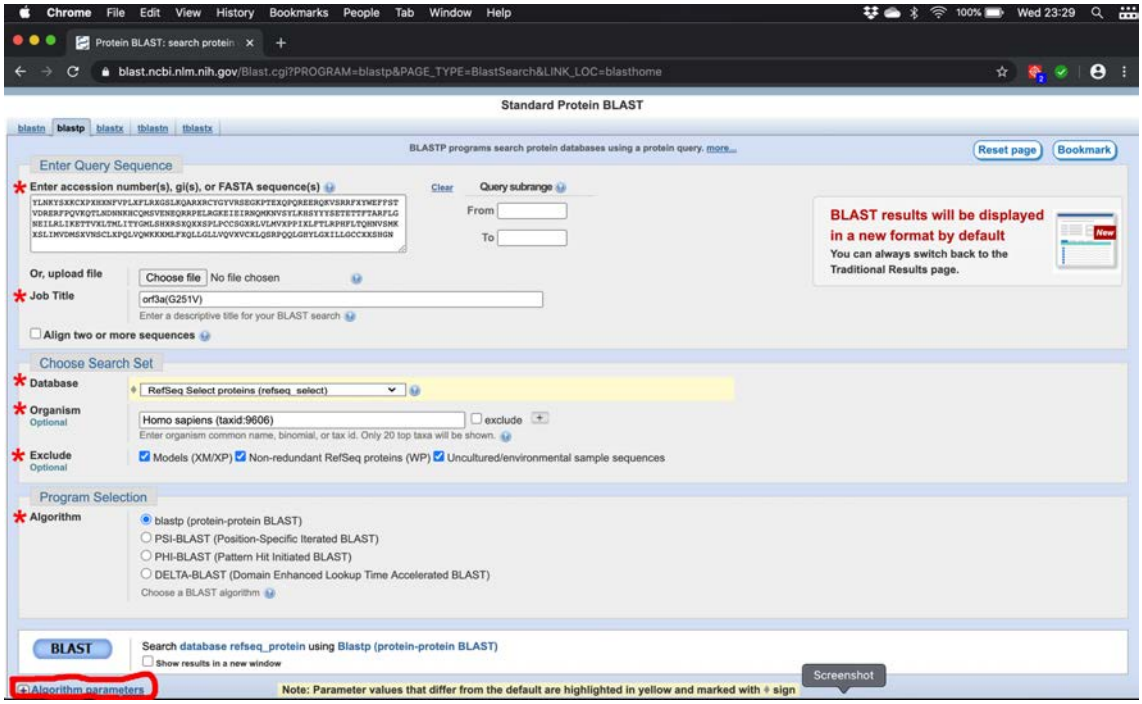
AS53CN = Antisense peptide (5'-3' mRNA reading) C-terminus to N-terminus orientation

(5b) Separate BLAST searches should be carried out for each antisense sequence (AS35NC, AS35CN, AS53NC & AS53CN) using the https://blast.ncbi.nlm.nih.gov/Blast.cgi website to run a Protein BLAST:

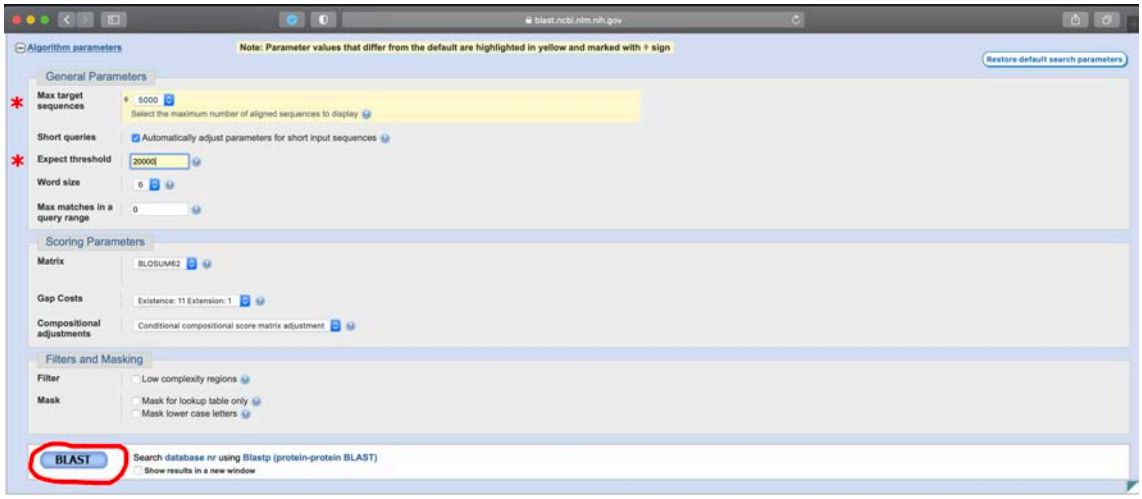


Under the header "Enter accession number, gi or FASTA sequence" paste in the antisense peptide sequence to be screened. Also enter the protein name into the Job title box.

Under the "Choose Search Set", "Database" select RefSeq Select proteins (refseq_select), under the "Organism" type in homo sapiens and select "Homo sapiens (taxid:9606)", also select the tick boxes for Exclude "Models (XM/XP)", "Non-redundant RefSeq proteins (WP)" and "Uncultured/environmental sample sequences" plus select Algorithm "blastp (protein-protein BLAST)" – see red * marks below:




Next click on "Algorithm parameters" (circled in red) which will go to:



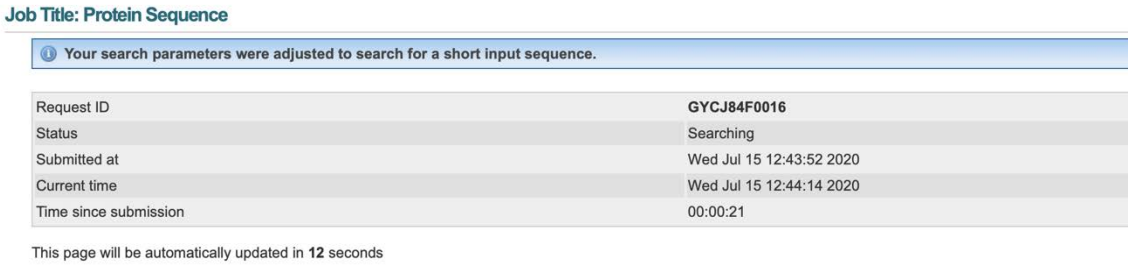

Change "Max target sequences" to 5,000 from the dropdown and "Expect threshold" to 20,000 – leave all other settings as the defaults (the Short

queries box "Automatically adjust parameters for short input sequences" is checked). The default Matrix under Scoring Parameters is BLOSUM62 (for information about this and other options have a look at Pearson 2013). Then click on BLAST (circled in red) and wait for results to appear.

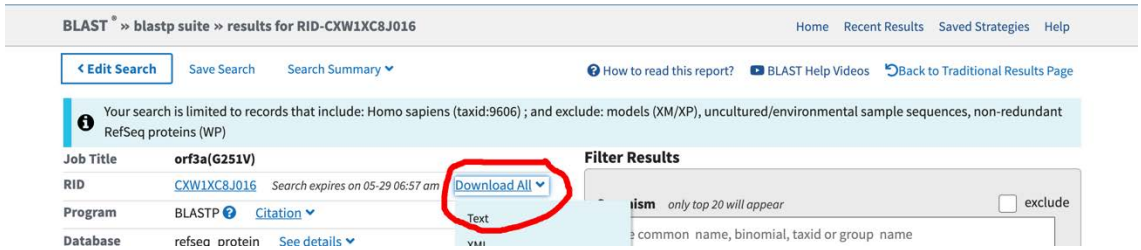
Note that any sequence with 30 or less amino acids will be automatically searched with adjusted parameters. The short sequence parameters automatically changed for short sequences like this will be the Expected threshold (which will be increased to 200,000), the Word size (which will reduce to 2), the Matrix (which will change to PAM30) and the Compositional adjustments (which will be set to no adjustment).

(5c) After clicking BLAST the screen will initially show:


If the sequence searched comprises 30 or less amino acids the following warning will then show:




After the search has finished the screen will show the following. Click on Download All dropdown and select Text. Save txt file to an appropriate folder (this file will be used in sections 6 (pages 20-30) and 7 (pages 31-39). A screenshot or save as pdfs file should be created after selecting each of the Descriptions, Graphic Summary and Alignment tabs. Saving as pdf's preserves the links that can be used in data analysis more easily later, for an example and other files that can be downloaded see section 6 (pages 20-30):



# 6: BLAST data extraction

(6a) Data should be extracted from each of the BLAST searches carried out, giving a total of four sets of data for each target protein (corresponding to the BLAST searches for AS35NC, AS35CN, AS53NC and AS53CN.

Since the AS35NC and AS53NC peptides will theoretically bind to the SNC peptide this suggests that anything similar to the AS35NC or AS53NC peptides would bind to the SNC peptide. In the BLAST search Alignments, the Query corresponds to the protein searched (AS35NC or AS53NC peptide) and the Sbjct corresponds to the named protein that is similar. The Query residue numbers are identical to the SNC numbers so for example if the query numbers were 4-9 and the Sbjct residues were 321-326 this would suggest that Sbjct protein 321-326 theoretically binds residues 4-9 of the SNC protein.

(6b) The situation with the AS35CN and AS53CN peptides is similar in that they will theoretically bind to the SCN peptide and this suggests that anything similar to the AS35CN or AS53CN peptides would bind to the SCN peptide. The SCN peptide corresponds to the SNC peptide in reverse. If the SNC was a 14 amino acid peptide the N-C direction numbering would be 1-14, therefore the C-N numbering would be 14-1. Blast searches always number sequences 1-14, which means the results for AS35CN and AS53CN peptides searched using BLAST will always be numbered in the wrong direction and the numbers will need to be converted.

The easiest way to do this is to create a table with the numbers from start to finish of the sequence in a column in ascending order, in this case 1-14, and a second column with the numbers in descending order – starting from the total number of amino acids of the SNC protein, in this case 14-1:

| | BLAST Query numbering AS35CN/AS53CN | NC numbering |
|---|---|---|
| 1st residue | 1 | 14 |
| | 2 | 13 |
| | 3 | 12 |
| | 4 | 11 |
| | 5 | 10 |
| | 6 | 9 |
| | 7 | 8 |
| | 8 | 7 |
| | 9 | 6 |
| | 10 | 5 |
| | 11 | 4 |
| | 12 | 3 |
| | 13 | 2 |
| Last residue | 14 | 1 |

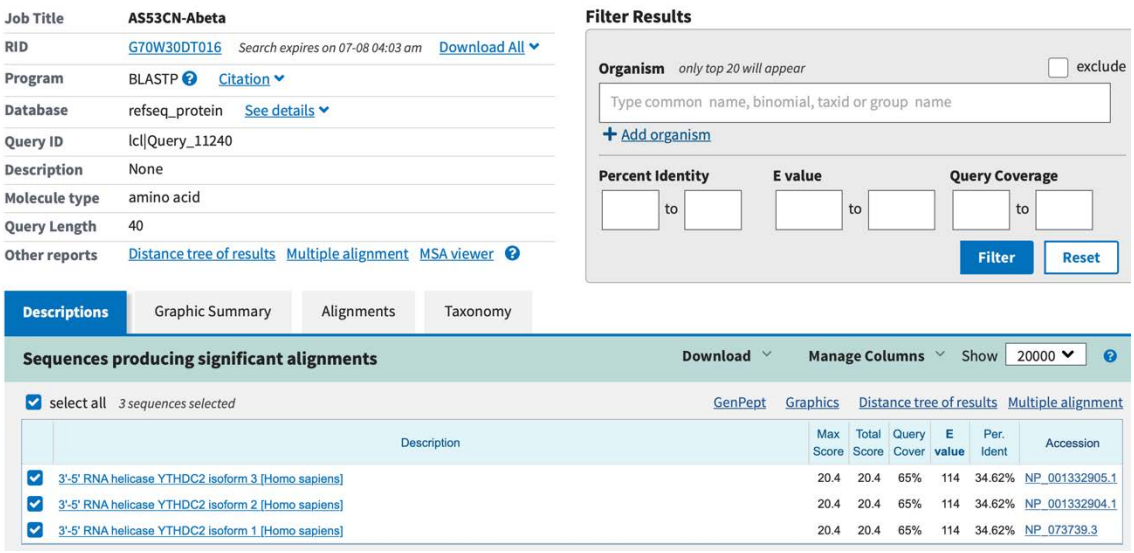
In the BLAST search Alignments, the Query corresponds to the protein searched and the Sbjct corresponds to the named protein that is similar. If an alignment were found where the Query region on the BLAST search results from an AS35CN and AS53CN peptide was 4-9 (circled in red) using the table above the NC numbering would be 11-6 (circled in blue). The 11-6 region is therefore the 6-11 region of the SNC, which is the protein originally used to generate the antisense peptides.

(6c) For the purposes of illustrating how to analyse the data the Amyloid-ß 1-40 peptide (Aß) will be used as an example:

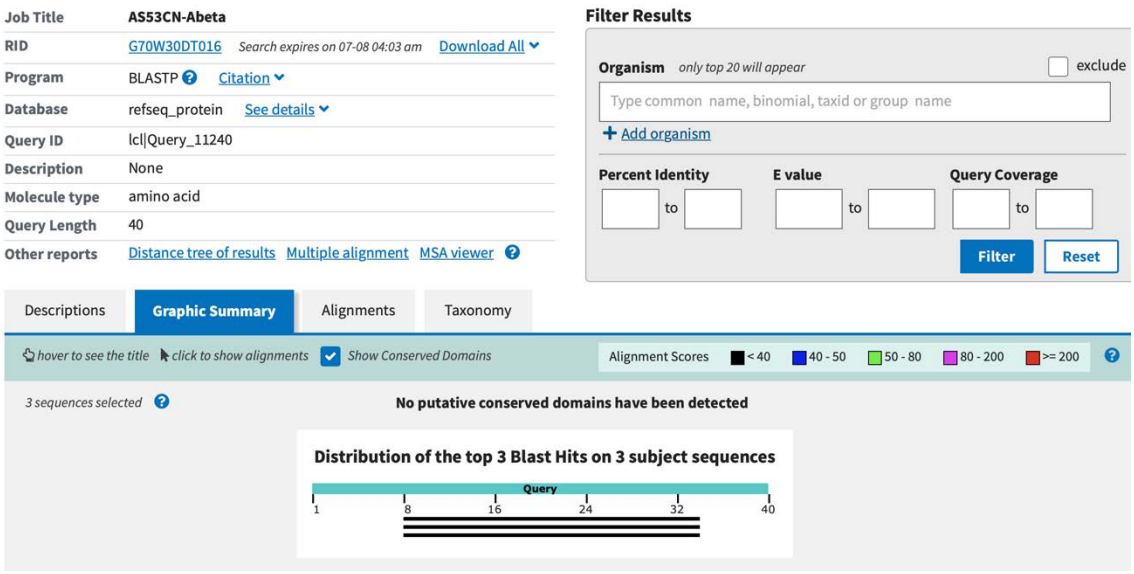
| | |
|---|---|
| Coding mRNA for Aß | GATGCAGAATTCCGACATGACTCAGGATATGAAGTTCATCAT CAAAAATTGGTGTTCTTTGCAGAAGATGTGGGTTCAAACAAA GGTGCAATCATTGGACTCATGGTGGGCGGTGTTGTC |
| SNC-Aß | DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV |
| SCN-Aß | VVGGVMLGIIAGKNSGVDEAFFVLKQHHVEYGSDHRFEAD |
| AS35NC-Aß | LRLKAVLSPILQVVVFNHKKRLLHPSLFPRXXPEYHPPQQ |
| AS35CN-Aß | QQPPHYEPXXRPFLSPHLLRKKHNFVVVQLIPSLVAKLRL |
| AS53NC-Aß | ICFESMVXSIFNMMLFQHEKCFIHTXVFTCDNSEHHATND |
| AS53CN-Aß | DNTAHHESNDCTFVXTHIFCKEHQFLMMNFISXVMSEFCI |

(6d) Running a BLAST search, as described in section 5 above (pages 17-19) for the AS53CN-Aß peptide gave the set of results:

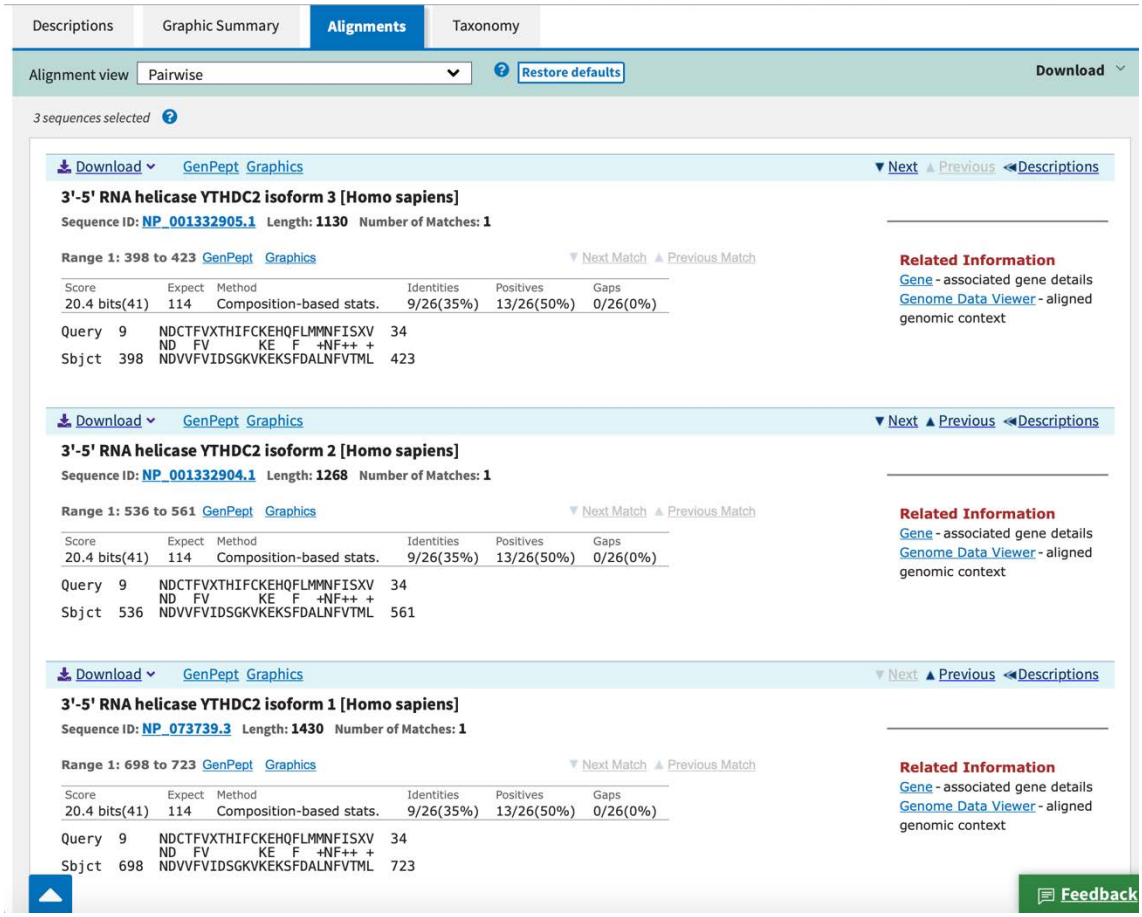
(i) Descriptions:



## (ii) Graphic Summary:




## (iii) Alignments:



(iv) Text File, with the key elements that should be extracted highlighted in red are best put into a table format and saved:

```
RID: G70W30DT016
Job Title:AS53CN-Abeta
Program: BLASTP
Database: refseq_protein NCBI Protein Reference Sequences
Query #1:  Query ID: lcl|Query_11240 Length: 40

Sequences producing significant alignments:

                                                                    Max    Total Query   E    Per.
Description                                                         Score  Score cover Value Ident  Accession
3'-5' RNA helicase YTHDC2 isoform 3 [Homo sapiens]                  20.4   20.4  65%   114   34.62  NP_001332905.1
3'-5' RNA helicase YTHDC2 isoform 2 [Homo sapiens]                  20.4   20.4  65%   114   34.62  NP_001332904.1
3'-5' RNA helicase YTHDC2 isoform 1 [Homo sapiens]                  20.4   20.4  65%   114   34.62  NP_073739.3


Alignments:
>3'-5' RNA helicase YTHDC2 isoform 3 [Homo sapiens]
Sequence ID: NP_001332905.1 Length: 1130
Range 1: 398 to 423

Score:20.4 bits(41), Expect:114,
Method:Composition-based stats.,
Identities:9/26(35%), Positives:13/26(50%), Gaps:0/26(0%)

Query  9    NDCTFVXTHIFCKEHQFLMMNFISXV  34
            ND  FV      KE  F  +NF++ +
Sbjct  398  NDVVFVIDSGKVKEKSFDALNFVTML  423


>3'-5' RNA helicase YTHDC2 isoform 2 [Homo sapiens]
Sequence ID: NP_001332904.1 Length: 1268
Range 1: 536 to 561

Score:20.4 bits(41), Expect:114,
Method:Composition-based stats.,
Identities:9/26(35%), Positives:13/26(50%), Gaps:0/26(0%)

Query  9    NDCTFVXTHIFCKEHQFLMMNFISXV  34
            ND  FV      KE  F  +NF++ +
Sbjct  536  NDVVFVIDSGKVKEKSFDALNFVTML  561


>3'-5' RNA helicase YTHDC2 isoform 1 [Homo sapiens]
Sequence ID: NP_073739.3 Length: 1430
Range 1: 698 to 723

Score:20.4 bits(41), Expect:114,
Method:Composition-based stats.,
Identities:9/26(35%), Positives:13/26(50%), Gaps:0/26(0%)

Query  9    NDCTFVXTHIFCKEHQFLMMNFISXV  34
            ND  FV      KE  F  +NF++ +
Sbjct  698  NDVVFVIDSGKVKEKSFDALNFVTML  723
```

(6e) **Any results with a % gaps greater than 0 should be reviewed as these may not be compatible with protein-protein binding.**

Gaps are where a space is introduced between two amino acids in either the Query or Sbjct sequence to achieve better alignment, they are indicated with one or more "-" in between two amino acid residues (circled in red).

```
>melanoma-associated antigen 10 [Homo sapiens]
Sequence ID: NP_066386.3 Length: 369
Range 1: 170 to 199

Score:18.5 bits(36), Expect:12367,
Method:Compositional matrix adjust.,
Identities:11/43(26%), Positives:16/43(37%), Gaps:17/43(39%)

Query  22   DHYP----KDKECTXTEDVKRIRLSLSKNIYXFNXKETDNTGH  60
            DH+P    +  EC              ++  + KE D TGH
Sbjct  170  DHFPLLFSEASECMLL----------VFGIDVKEVDPTGH  199
```

The key is that gaps are not present in protein sequences but are artifacts of the BLAST search methods. This means that gaps could create an artificial alignment that in reality would not lead to a protein-protein interaction. The higher the % gaps the more likely the result would not be compatible with binding.

The length of ungapped alignment within the BLAST results needs to be sufficient to suggest possible binding and as such needs to be accounted for when taking results with gaps forward. In this example the selection circled in red could still potentially be used:

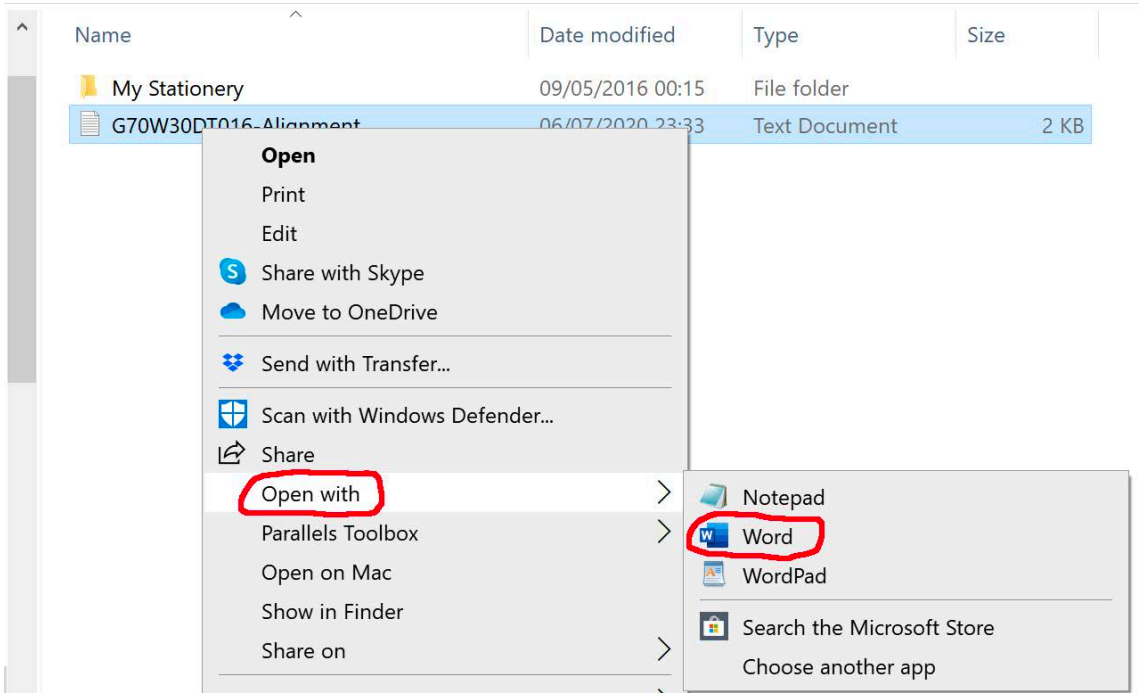
```
>peptidyl-prolyl cis-trans isomerase-like 4 [Homo sapiens]
Sequence ID: NP_624311.1 Length: 492
Range 1: 126 to 156

Score:19.2 bits(38), Expect:7508,
Method:Compositional matrix adjust.,
Identities:8/32(25%), Positives:18/32(56%), Gaps:1/32(3%)

Query  173  FXYINKSLSLRIRKVDETXTCEKDKPYHDLTV  204
            F  + + + + I+K++ET   +   PY D+ +
Sbjct  126  FGEVTEGMDI-IKKINETFVDKDFVPYQDIRI  156
```

The resultant alignment would be used along with recalculated % identity and % positives for the shorter segment. The E value and the statistical analysis would no longer be valid.

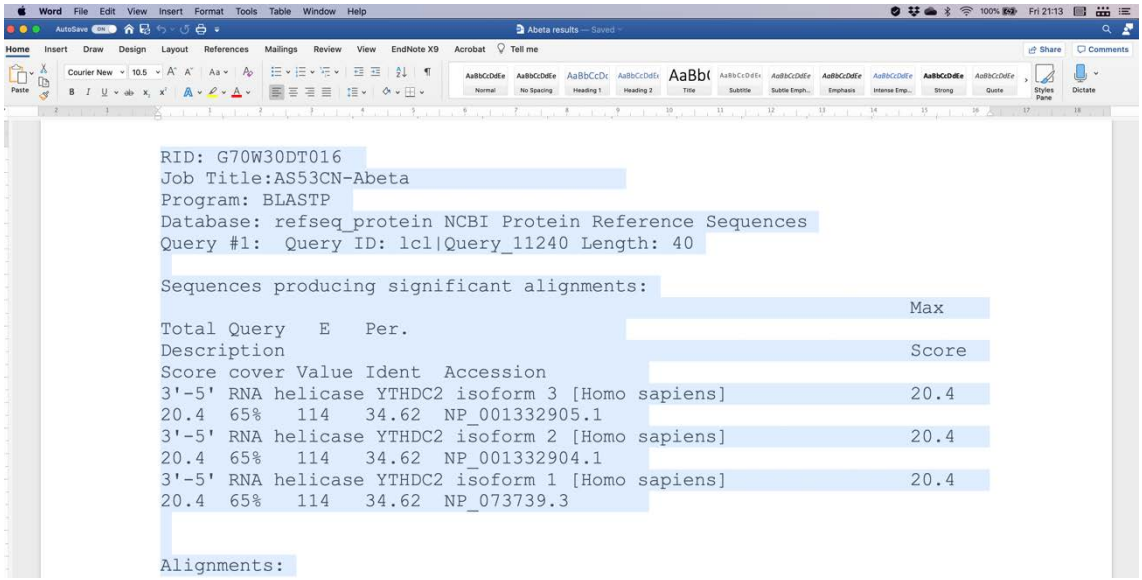
```
>peptidyl-prolyl cis-trans isomerase-like 4 [Homo sapiens]
Sequence ID: NP_624311.1 Length: 492
Range 1: 136 to 156

Score:19.2 bits(38), Expect:7508,
Method:Compositional matrix adjust.,
Identities:7/21(33%), Positives:13/21(62%), Gaps:0/21(0%)

Query  184  IRKVDETXTCEKDKPYHDLTV  204
            I+K++ET   +   PY D+ +
Sbjct  136  IKKINETFVDKDFVPYQDIRI  156
```

**For the purpose of antisense binding 0% gaps is best, and generally the majority of sequences with >0% gaps are discarded.**

(6f) The easiest method to extract the data is to open the txt file in Word, select the txt file and then right-click to bring up the options. Select Open With (circled in red) and then select Word (circled in red), just click OK for any conversion settings are suggested and then save document as Protein Name Results.docx. Keep the original txt file as it contains the aligned sequences that are needed in section 7c (page 32 below) and is a record of the results of the BLAST search:



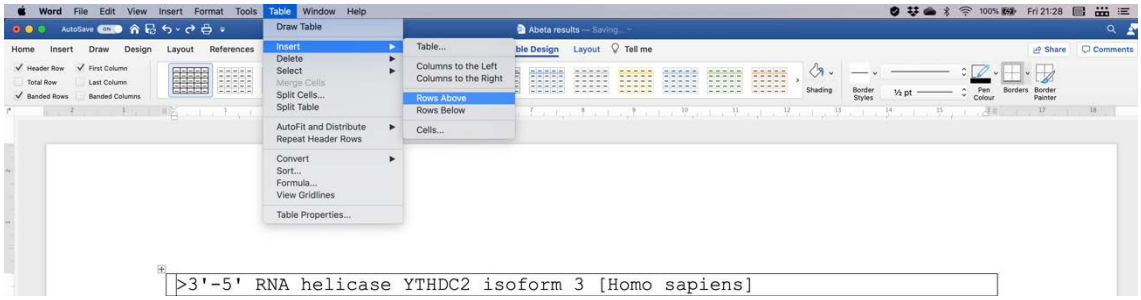

Once saved delete the initial parts of the file down to alignments (highlighted in blue):



```
RID: G70W30DT016
Job Title:AS53CN-Abeta
Program: BLASTP
Database: refseq_protein NCBI Protein Reference Sequences
Query #1:  Query ID: lcl|Query_11240 Length: 40

Sequences producing significant alignments:
                                                                    Max
Total Query   E   Per.
Description                                                         Score
Score cover Value Ident  Accession
3'-5' RNA helicase YTHDC2 isoform 3 [Homo sapiens]                  20.4
20.4  65%   114   34.62  NP_001332905.1
3'-5' RNA helicase YTHDC2 isoform 2 [Homo sapiens]                  20.4
20.4  65%   114   34.62  NP_001332904.1
3'-5' RNA helicase YTHDC2 isoform 1 [Homo sapiens]                  20.4
20.4  65%   114   34.62  NP_073739.3


Alignments:
```

Then select complete text and using the Table option convert text to table (highlighted in blue), make sure converts to one column table:


Then add a row at the top and type in Name of protein similar to AS53CN-protein name in that row. Then select add columns to the right and add a total of seven columns to give an eight-column table. Then put the headings in table below:




In the table delete the > symbol at the start of the protein name. Then copy the NP number (circled in red) and paste in column 2; the length number (circled in red) and paste in column 3; the range numbers (circled in red) and paste in column 4; the Query start and end numbers (circled in red) and paste in column 5; the % ID (circled in red) and paste in column 6; the % +ve (circled in red) and paste in column 7; and finally, the % Gaps (circled in red) and paste in column 8. This will give a table like this:

| Name of protein similar to AS53CN-Aß | Protein ID | Size | Residues | Query residues | % ID | % +ve | % Gaps |
|---|---|---|---|---|---|---|---|
| 3'-5' RNA helicase YTHDC2 isoform 3 [Homo sapiens] | NP_001332905.1 | 1130 | 398-423 | 9-34 | 35 | 50 | 0 |
| Sequence ID: NP_001332905.1 Length: 1130 | | | | | | | |
| Range 1: 398 to 423 | | | | | | | |
| | | | | | | | |
| Score:20.4 bits(41), Expect:114, | | | | | | | |
| Method:Composition-based stats., | | | | | | | |
| Identities:9/26(35%), Positives:13/26(50%), | | | | | | | |
| Gaps:0/26(0%) | | | | | | | |
| | | | | | | | |
| Query 9 NDCTFVXTHIFCKEHQFLMMNFISXV 34 | | | | | | | |
| ND FV KE F +NF++ + | | | | | | | |
| Sbjct 398 NDVVFVIDSGKVKEKSFDALNFVTML 423 | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| >3'-5' RNA helicase YTHDC2 isoform 2 [Homo | | | | | | | |

Delete the highlighted rows in blue above, down to the next protein name and repeat the same process above, until all of the proteins have been processed to give a table like this:

| Name of protein similar to AS53CN-Aß | Protein ID | Size | Residues | Query residues | % ID | % +ve | % Gaps |
|---|---|---|---|---|---|---|---|
| 3'-5' RNA helicase YTHDC2 isoform 1 | NP_073739.3 | 1430 | 698-723 | 9-34 | 35 | 50 | 0 |
| 3'-5' RNA helicase YTHDC2 isoform 2 | NP_001332904.1 | 1268 | 536-561 | 9-34 | 35 | 50 | 0 |
| 3'-5' RNA helicase YTHDC2 isoform 3 | NP_001332905.1 | 1130 | 398-423 | 9-34 | 35 | 50 | 0 |

(6g) Make sure that none of the proteins have a % Gaps > 0 as this data should have been discarded in section 6e above (page 23), if there are any rows with a % Gaps > 0 they should be reviewed as described.

(6h) If the antisense peptide sequence screened is an AS35NC or AS53NC the Query residue numbers correspond directly to the SNC residue numbering (see 6a, page 20 above). However, in this case the AS53CN peptide was used, which means the numbers need to be converted as described in 6b above (pages 20-21). Using a table with 1-40 ascending in Column A and 40-1 descending in Column B (40 is the number of residues in the SNC-Aß [see Table 6c, page 21 above], can be determined using word count for the sequence):

| CN - Residue | NC - Residue |
|---|---|
| 1 | 40 |
| 2 | 39 |
| 3 | 38 |
| 4 | 37 |
| 5 | 36 |
| 6 | 35 |
| 7 | 34 |
| 8 | 33 |
| 9 | 32 |
| 10 | 31 |
| 11 | 30 |
| 12 | 29 |
| 13 | 28 |
| 14 | 27 |
| 15 | 26 |
| 16 | 25 |
| 17 | 24 |
| 18 | 23 |
| 19 | 22 |
| 20 | 21 |
| 21 | 20 |
| 22 | 19 |
| 23 | 18 |
| 24 | 17 |
| 25 | 16 |
| 26 | 15 |
| 27 | 14 |
| 28 | 13 |
| 29 | 12 |
| 30 | 11 |
| 31 | 10 |
| 32 | 9 |
| 33 | 8 |
| 34 | 7 |
| 35 | 6 |
| 36 | 5 |
| 37 | 4 |
| 38 | 3 |
| 39 | 2 |
| 40 | 1 |

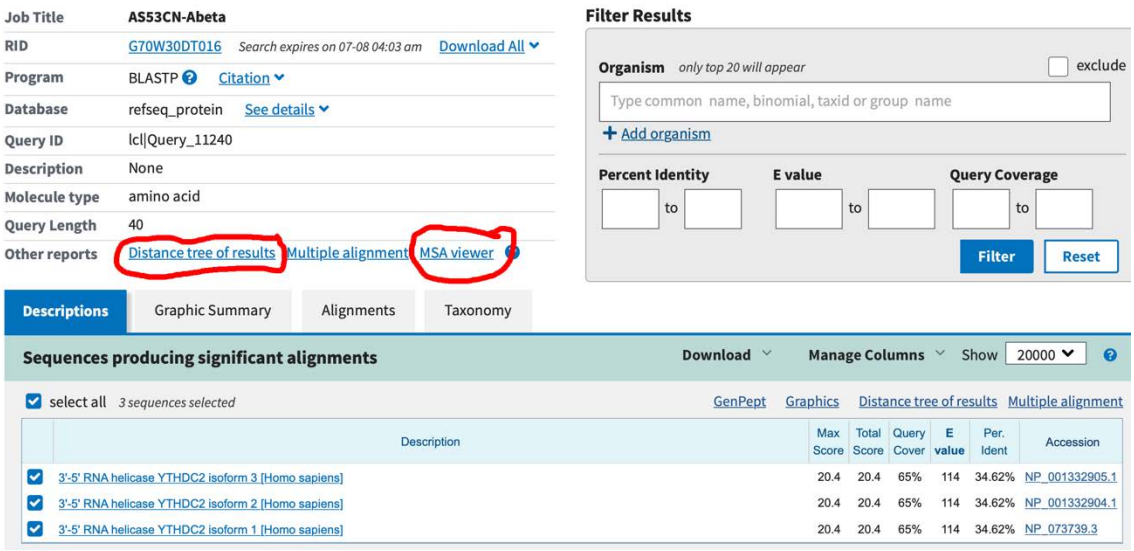
The 9-34 Query residues (circled in red) correspond to 32-7 (circled in blue). The Query residues with the revised numbering corresponding to the Aß residues 32-7 and the results table can be modified to show this:

Table 6h

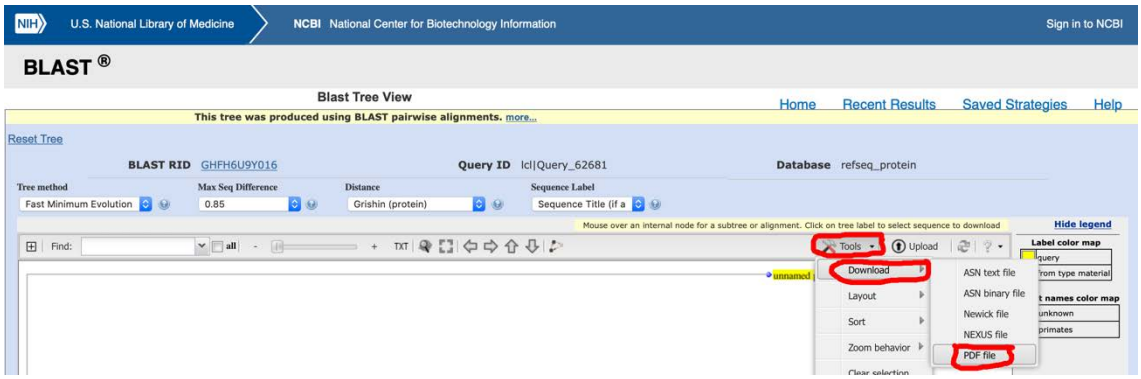
| Human Protein that theoretically binds human Aß | Protein ID | Size | Residues | Aß residues | % ID | % +ve | % Gaps |
|---|---|---|---|---|---|---|---|
| 3'-5' RNA helicase YTHDC2 isoform 1 | NP_073739.3 | 1430 | 698-723 | 32-7 | 35 | 50 | 0 |
| 3'-5' RNA helicase YTHDC2 isoform 2 | NP_001332904.1 | 1268 | 536-561 | 32-7 | 35 | 50 | 0 |
| 3'-5' RNA helicase YTHDC2 isoform 3 | NP_001332905.1 | 1130 | 398-423 | 32-7 | 35 | 50 | 0 |

Thus 3'-5' RNA helicase YTHDC2 isoform 1 residues 698-723 theoretically binds to Aß residues 7-32.

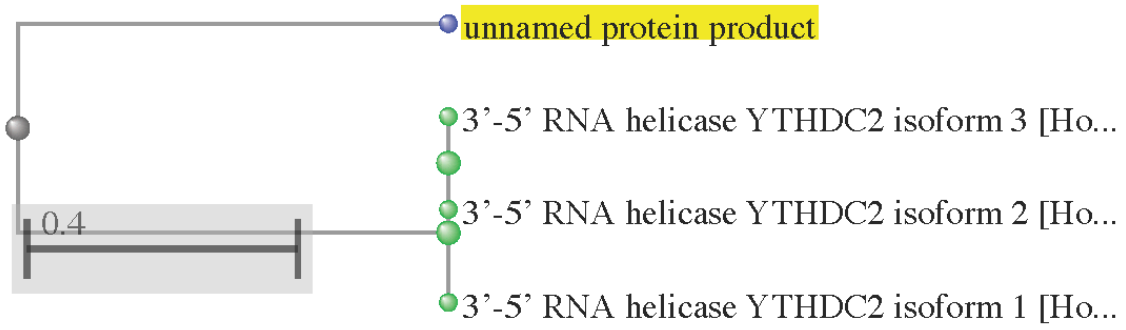
(6i) Within the results files there are other potentially useful pieces of information that can be used when writing up the results. There are also other useful sets of data that can be downloaded from the Blast search while online:



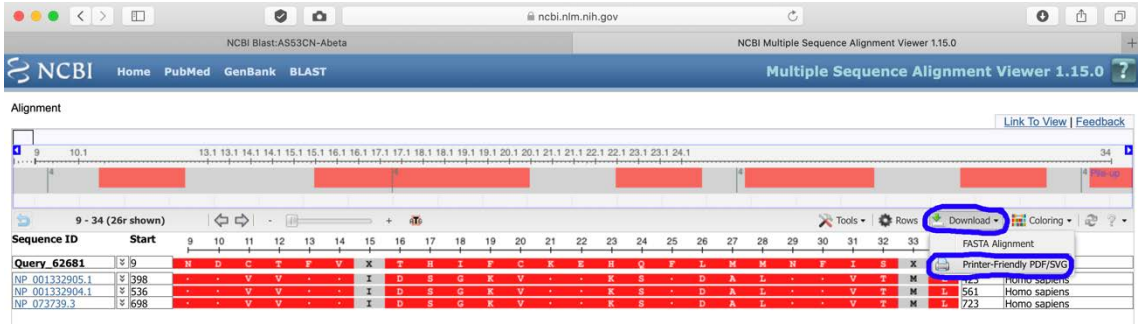

The distance tree of results (click link circled in red) can be useful and also the MSA viewer (again click link circled in red). From these links the files can be downloaded as pdfs. For the Distance tree of results: after clicking on link go to Tools (circled in red), then Download (circled in red), then select PDF:



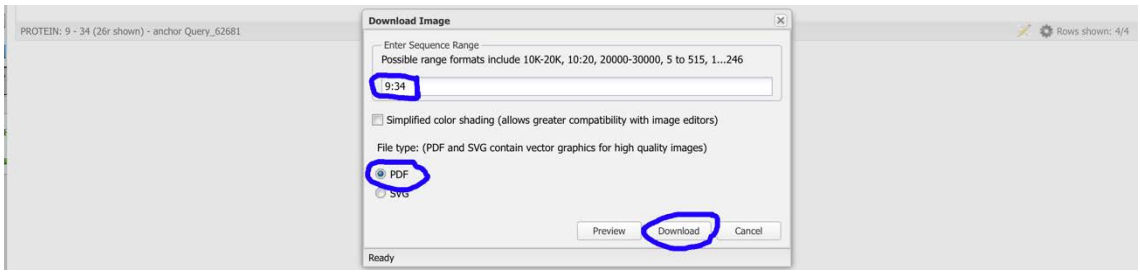

This will give the following where the unnamed protein product is the AS53CN-Aß and the results indicate how closely related the sequences are:



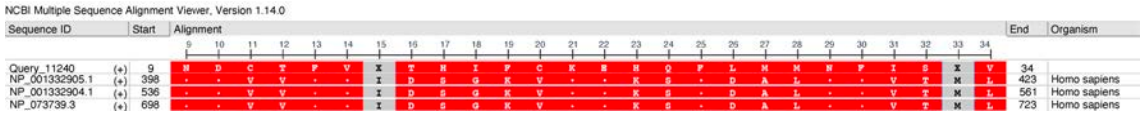

Multiple sequence alignment (from MSA viewer) results: after clicking on link go to Download (circled in blue), then select Printer-Friendly PDF/SVG, which will bring up a second box:



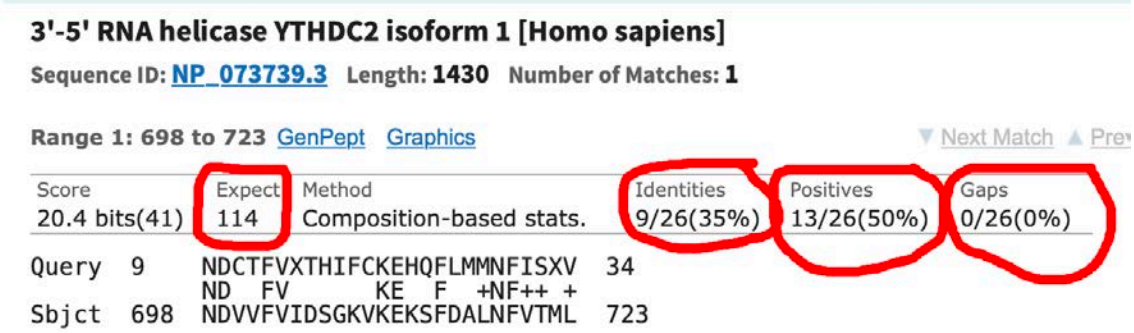

In the second box the Possible range formats box (circled in blue) should have the 9-34 for the fragment of AS53CN-Aß that is aligned, if there are other alignments may have a range covering all of them. Select PDF (circled in blue) and click Download (circled in blue):



This will give the following where the query is AS53CN-Aß and the alignments are for the proteins from table at end of 6f above (page 26), note in this case the numbering is AS53CN-Aß sequence numbering and not the corrected Aß SNC sense numbering:




(6j) Within the saved Alignments files there are a number of components that are also useful: NEW TABLE HERE – Alignments, Expect, plus bits from 6h




(6k) The Expect (E-value) is a statistical indication of similarity between the Query and Sbjct, the smaller the E-value, the better the match. Within the extracted data files created in above the Identities, Positives and Gaps data has been extracted.

(i) The Identities is the % of amino acids in the Query sequence that are identical to those in the Sbjct sequence – these are shown in the middle row between the Query and Sbjct rows as letters.

(ii) The number of Positives includes the identical amino acids. In the context of alignments displayed in the BLAST output, the Positives are those non-identical substitutions that receive a Positives score in the underlying scoring matrix, BLOSUM62 by default. Most often, Positives indicate a conservative substitution or substitutions that are often observed in related proteins. In effect the Positives is the % of amino acids in the Query sequence that are similar to those in the Sbjct sequence – these are shown in the middle row between the Query and Sbjct rows as a + symbol. Similarity of amino acids in this case is often based on similar structural features in the R group of the amino acids.

(iii) Gaps are where a space is introduced between two amino acids in either the Query or Sbjct sequence to achieve better alignment, for the purpose of antisense binding a 0% Gaps is essential, and sequence alignments with Gaps are normally discarded (see 6e above, page 23).

# 7: Molecular recognition analysis

(7a) The basis for the molecular recognition theory is that amino acids encoded by the sense strand of DNA will bind the corresponding amino acids encoded by the antisense strand. Using the BLAST search with a scoring matrix to identify similarities has limitations and may identify sequences that do not have a good potential to actually bind the target protein. The % identity from the BLAST search data represent identity with the antisense peptide and is much more useful than the % positives, which may give an inflated indication of potential binding.

The Molecular Recognition (MR) scoring is an alternative system that determines the potential of the target protein residues binding to the identified interacting protein residues based on antisense/sense interactions. The method determines the quantity of antisense/sense pairs within the identified target/interacting protein regions identified by the BLAST search. Using aligned target/interacting protein regions and the following table a score for the potential interaction is calculated. For example, if the first residue of the target is an A, then a score of 1 would be given if the first residue of the binding protein was either a C, G, R or S. If the first residue of the binding protein was any other amino acid, a score of 0 would be given. Then the process is repeated for each residue of the target/interacting protein regions:

(7b) Table for Molecular Recognition Scoring:

| Target Residue | Binding Protein Residue | MR Score |
|---|---|---|
| Ala (A) | Cys (C) Gly (G) Arg (R) Ser (S) | +1 |
| Cys (C) | Ala (A) Thr (T) | +1 |
| Asp (D) | Ile (I) Leu (L) Val (V) | +1 |
| Glu (E) | Phe (F) Leu (L) | +1 |
| Phe (F) | Glu (E) Lys (K) | +1 |
| Gly (G) | Ala (A) Pro (P) Ser (S) Thr (T) | +1 |
| His (H) | Met (M) Val (V) | +1 |
| Ile (I) | Asp (D) Asn (N) Tyr (Y) | +1 |
| Leu (L) | Asp (D) Glu (E) Lys (K) Asn (N) Gln (Q) | +1 |
| Lys (K) | Leu (L) Phe (F) | +1 |
| Met (M) | His (H) Tyr (Y) | +1 |
| Asn (N) | Ile (I) Leu (L) Val (V) | +1 |
| Pro (P) | Gly (G) Arg (R) Trp (W) | +1 |
| Gln (Q) | Leu (L) Val (V) | +1 |
| Arg (R) | Ala (A) Pro (P) Ser (S) Thr (T) | +1 |
| Ser (S) | Ala (A) Gly (G) Arg (R) Ser (S) Thr (T) | +1 |
| Thr (T) | Cys (C) Gly (G) Arg (R) Ser (S) Trp (W) | +1 |
| Val (V) | Asp (D) His (H) Asn (N) Gln (Q) Tyr (Y) | +1 |
| Trp (W) | Pro (P) Thr (T) | +1 |
| Tyr (Y) | Ile (I) Met (M) Val (V) | +1 |

(7c) For AS35NC and AS53NC sequences the Query residues from a BLAST search should align with the SNC residues for comparison. For AS35CN and AS53CN sequences the Query residues from a BLAST search should align with the SCN residues for comparison. The key is to use the Query residue numbers to identify the SNC or SCN residues to use.

Copy the alignments from the txt file (see section 6d, (iv) above, page 23) and paste into a word document (for ease always use a monospaced font such as Courier New). Save the document as a word file.

Open the Python results file for the target protein and then copy the SNC or SCN residues corresponding to the Query residues for each protein and paste then above the Query – see examples below. Then align the SNC or SNC residues with the Query residues. The SNC or SCN should then be aligned with the Sbjct residues.

Using Table 7b above (page 31) the Molecular Recognition (MR) score for the protein can be determined, using the SNC or SCN residues as the Target protein residues in the table and the Sbjct residues as the Binding Protein residues.

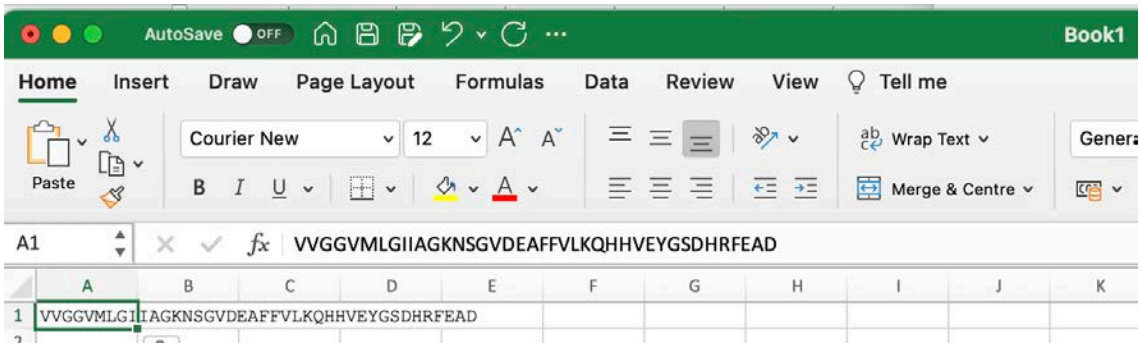
The total score is best expressed as a % of the total number of residues in the target sequence. The higher this % the more likely a binding interaction is likely to occur. This number will always be equal to or greater than the % identity score from the BLAST search.

(7d) The following efficient process of calculating the MR score for alignments obtained from antisense peptide Blast results using AS35CN, AS53CN, AS35NC, AS53NC employs Microsoft Excel to eliminate human error:
As an example, using data from the AS53CN-Aß Blast results below, plus the Aß SCN sequence detailed in section 6f (pages 24-27 above) the MR score will be determined for the BLAST alignment that predicted an interaction between Aß and the 3'-5' RNA helicase YTHDC2 isoform 1:

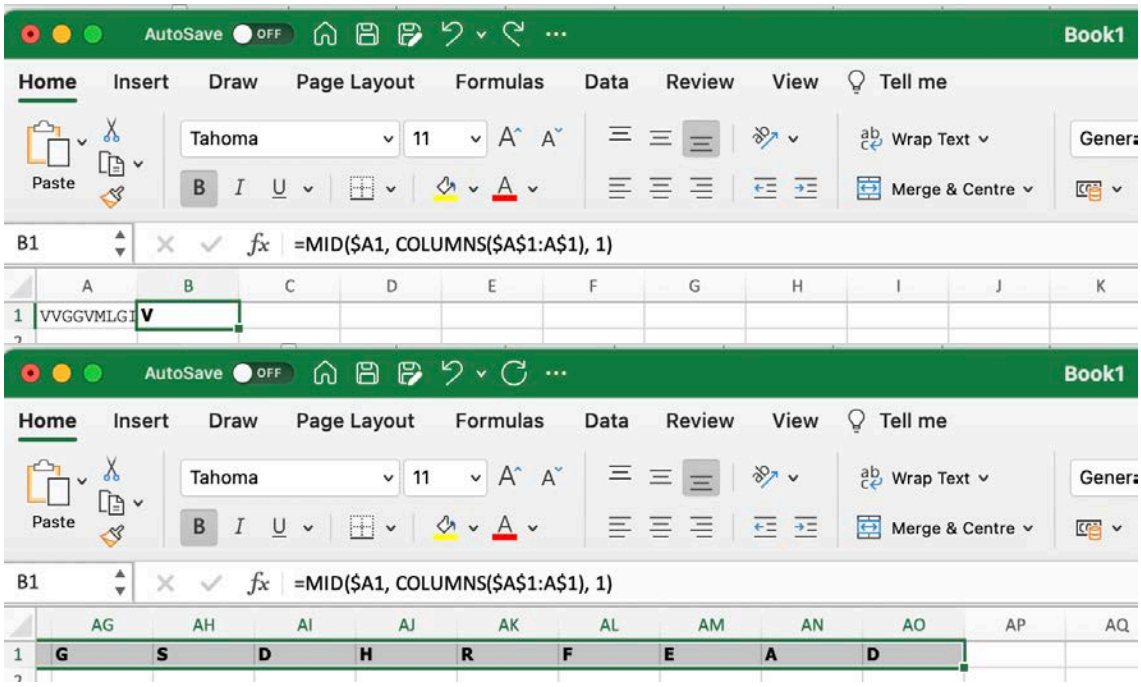
```
>3'-5' RNA helicase YTHDC2 isoform 1 [Homo sapiens]
Sequence ID: NP_073739.3 Length: 1430
Range 1: 698 to 723

Score:20.4 bits(41), Expect:114,
Method:Composition-based stats.,
Identities:9/26(35%), Positives:13/26(50%), Gaps:0/26(0%)

Query  9    NDCTFVXTHIFCKEHQFLMMNFISXV  34
            ND  FV      KE  F  +NF++ +
Sbjct  698  NDVVFVIDSGKVKEKSFDALNFVTML  723
```

(i) The whole SCN sequence was copied and pasted from the python outputs file into a single cell in an excel spreadsheet.

```
Input name: Aß
Input coding mRNA:
GATGCAGAATTCCGACATGACTCAGGATATGAAGTTCATCATCAAAAATTGGTGTTCTTTGC
AGAAGATGTGGGTTCAAACAAAGGTGCAATCATTGGACTCATGGTGGGCGGTGTTGTC

SNC - Aß = DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV

SCN - Aß = VVGGVMLGIIAGKNSGVDEAFFVLKQHHVEYGSDHRFEAD
```

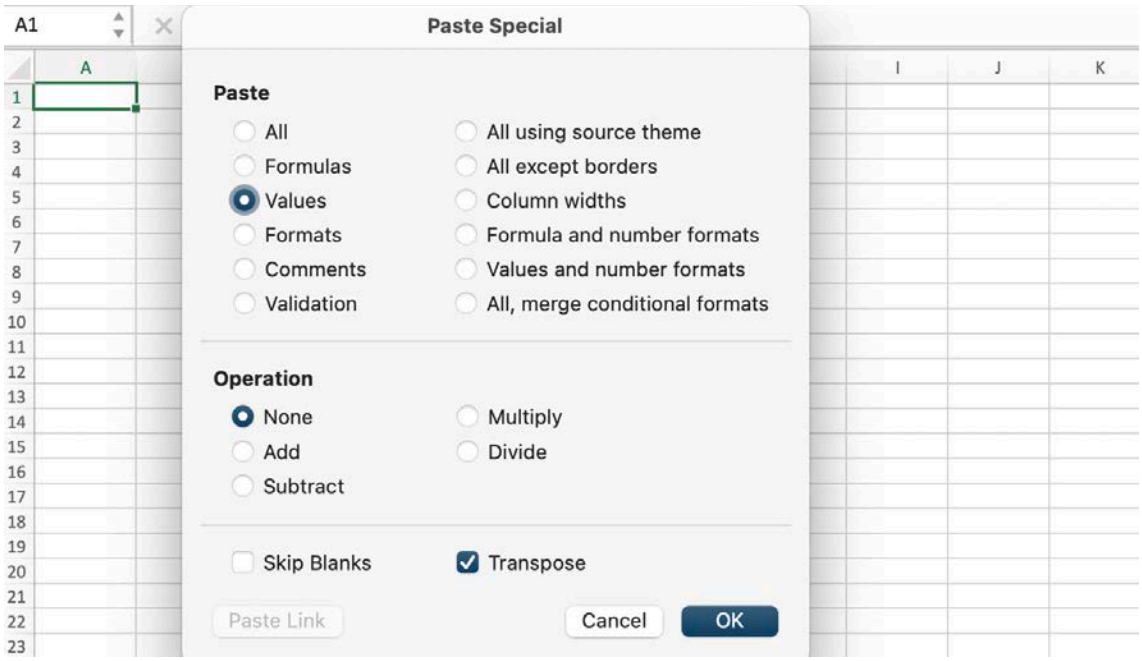


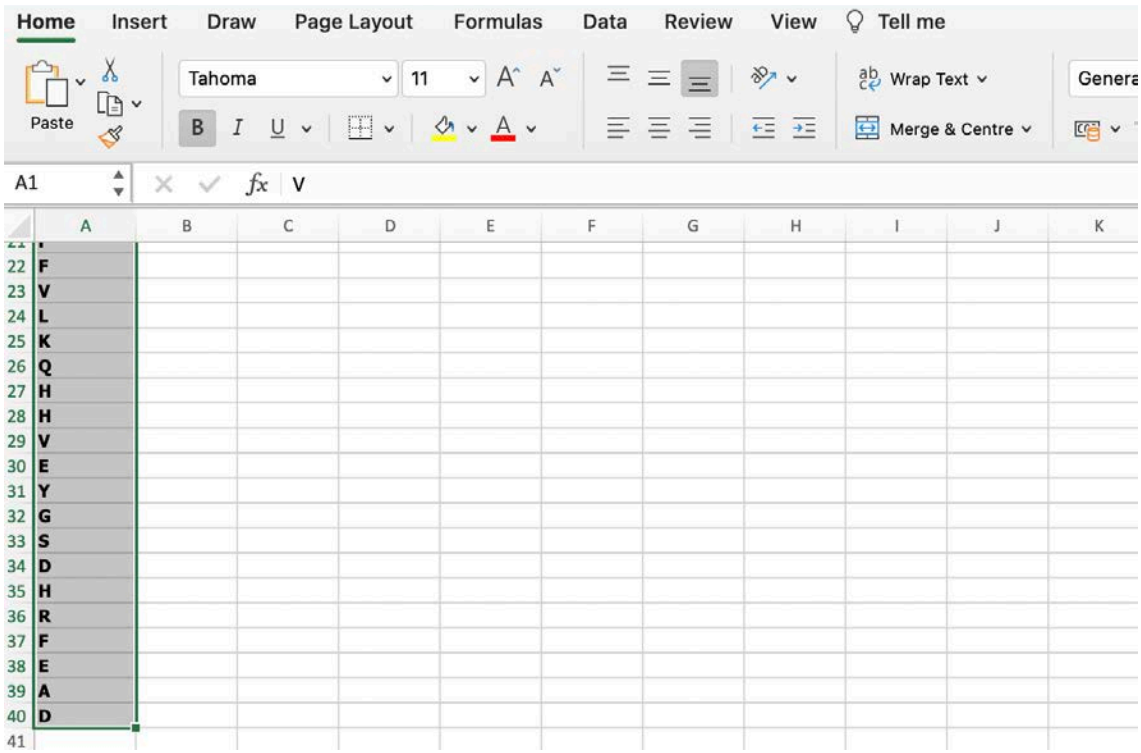

(ii) In cell B1, the following formula was then entered: **=MID($A1, COLUMNS($A$1:A$1), 1)**, where the first amino acid of the sequence appeared. To separate each amino acid of the sequence into one cell per column, cell B1 was dragged to the right across row one until the last amino acid had a position.




(iii) The amino acid sequenced was then copied and pasted in a separate sheet by using the Paste Special options 'Values' and 'Transpose'. By pasting the first amino acid in position A1 the rest amino acids were arranged to the positions

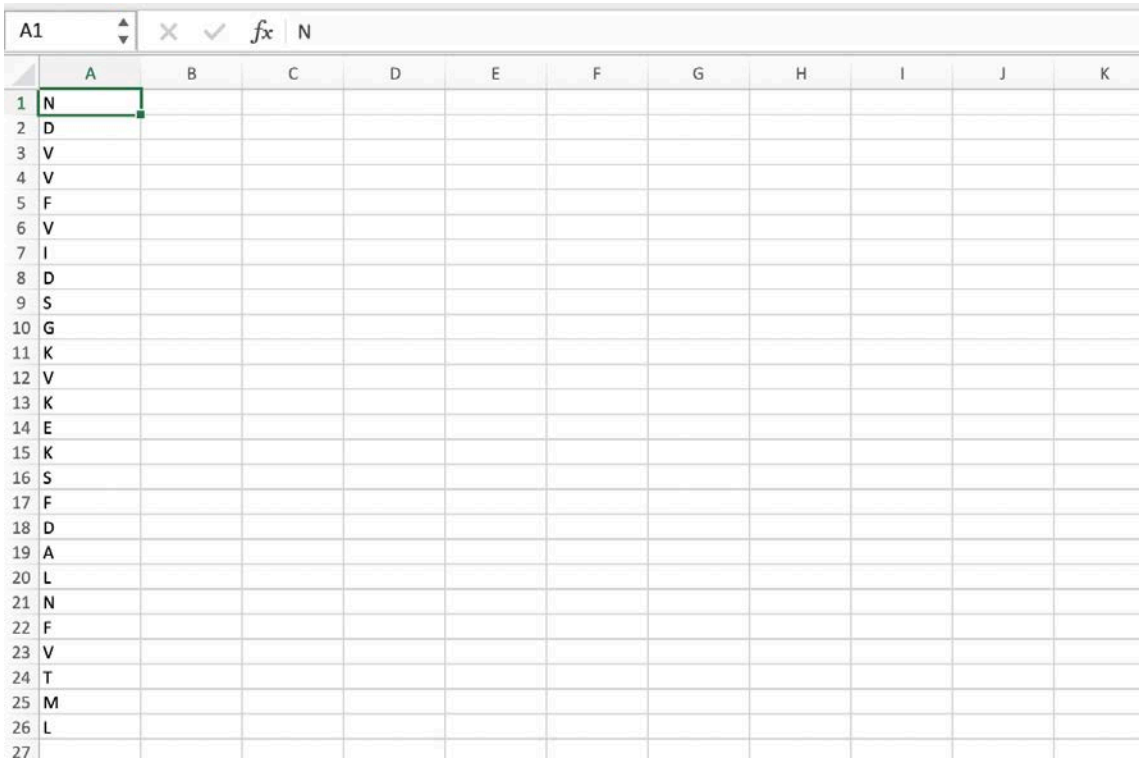
underneath so that the number of the row corresponded to the number of the amino acid residue.



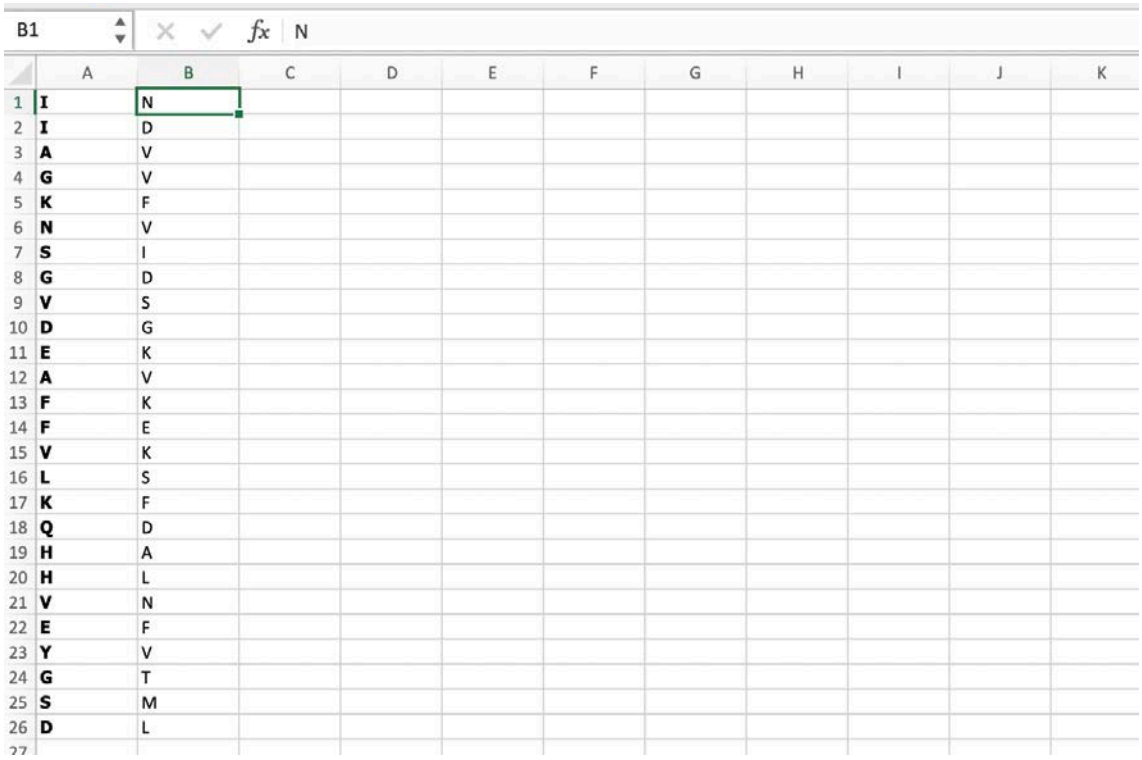




(iv) The interacting residue section of the SCN peptide was then isolated, in this example from the BLAST result detailed in section 7d above this was residues 9-34 (IIAGKNSGVDEAFFVLKQHHVEYGSD), pasted into a new excel file and saved as MR comparison.


(v) The residue section of the Sbjct sequence, in this case 3'-5' RNA helicase YTHDC2 isoform 1 residues 698-723 (NDVVFVIDSGKVKEKSFDALNFVTML) was also copied and pasted in an excel spreadsheet in a similar manner to the SCN sequence and steps (i) – (iii) above were repeated.

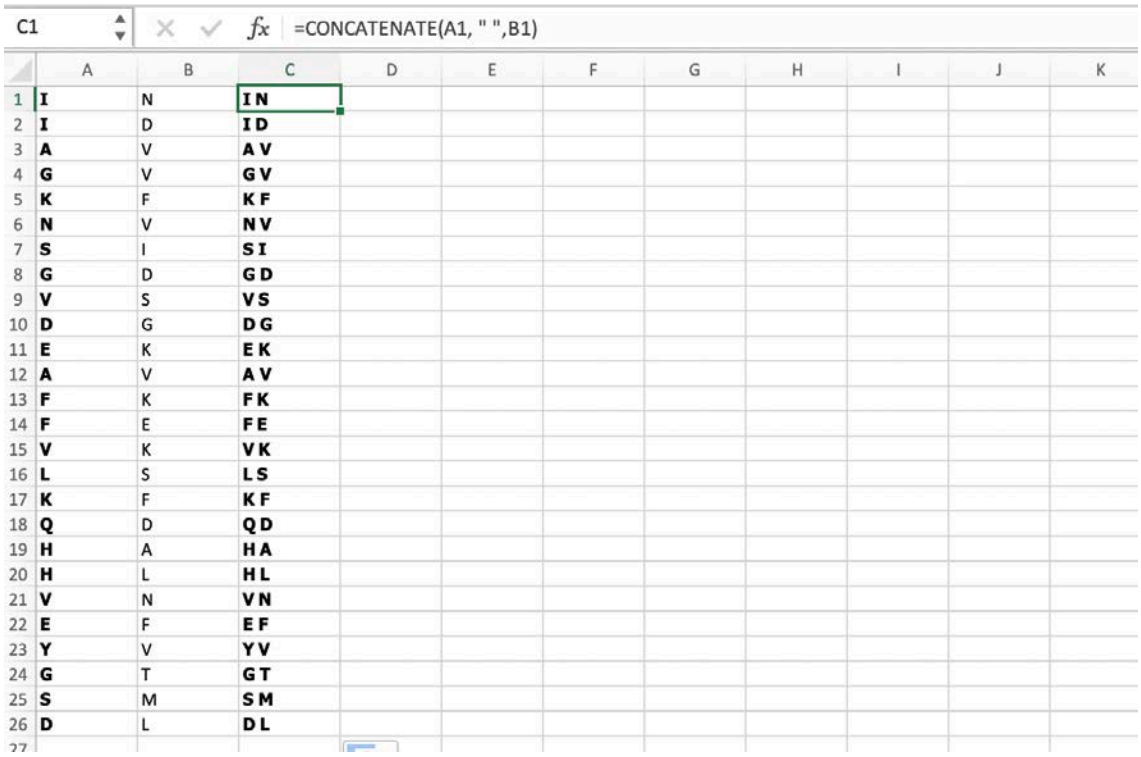



(vi) The above section was then copied and pasted next to the SCN sequence in the saved MR comparison excel file as below:



| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | I | N | | | | | | | | | |
| 2 | I | D | | | | | | | | | |
| 3 | A | V | | | | | | | | | |
| 4 | G | V | | | | | | | | | |
| 5 | K | F | | | | | | | | | |
| 6 | N | V | | | | | | | | | |
| 7 | S | I | | | | | | | | | |
| 8 | G | D | | | | | | | | | |
| 9 | V | S | | | | | | | | | |
| 10 | D | G | | | | | | | | | |
| 11 | E | K | | | | | | | | | |
| 12 | A | V | | | | | | | | | |
| 13 | F | K | | | | | | | | | |
| 14 | F | E | | | | | | | | | |
| 15 | V | K | | | | | | | | | |
| 16 | L | S | | | | | | | | | |
| 17 | K | F | | | | | | | | | |
| 18 | Q | D | | | | | | | | | |
| 19 | H | A | | | | | | | | | |
| 20 | H | L | | | | | | | | | |
| 21 | V | N | | | | | | | | | |
| 22 | E | F | | | | | | | | | |
| 23 | Y | V | | | | | | | | | |
| 24 | G | T | | | | | | | | | |
| 25 | S | M | | | | | | | | | |
| 26 | D | L | | | | | | | | | |

(vii) The following formula was subsequently copied and pasted in cell C1, to facilitate merging of columns A+B: **=CONCATENATE(A1, " ",B1)**. To complete merging of all the rows, cell C1 was dragged all the way down to last row containing an amino acid code (in this example row 26).

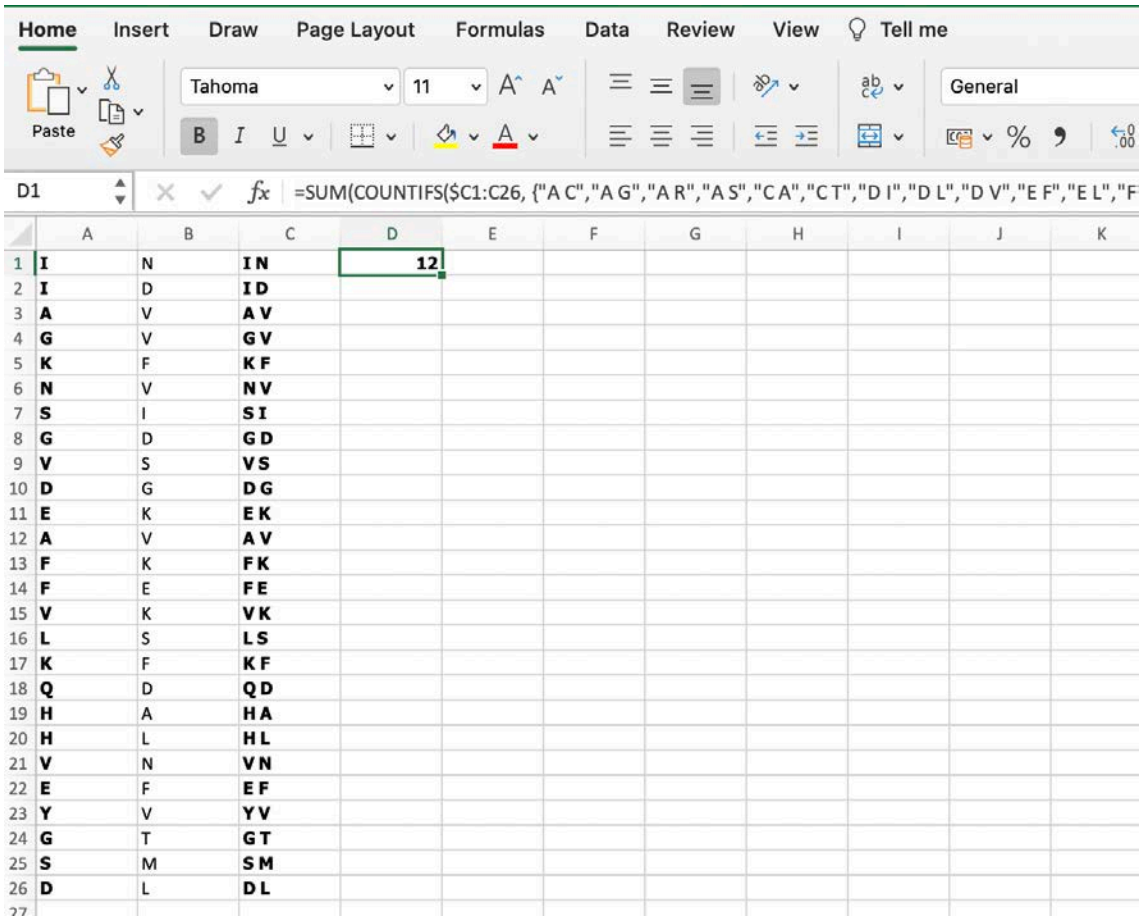


| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | I | N | I N | | | | | | | | |
| 2 | I | D | I D | | | | | | | | |
| 3 | A | V | A V | | | | | | | | |
| 4 | G | V | G V | | | | | | | | |
| 5 | K | F | K F | | | | | | | | |
| 6 | N | V | N V | | | | | | | | |
| 7 | S | I | S I | | | | | | | | |
| 8 | G | D | G D | | | | | | | | |
| 9 | V | S | V S | | | | | | | | |
| 10 | D | G | D G | | | | | | | | |
| 11 | E | K | E K | | | | | | | | |
| 12 | A | V | A V | | | | | | | | |
| 13 | F | K | F K | | | | | | | | |
| 14 | F | E | F E | | | | | | | | |
| 15 | V | K | V K | | | | | | | | |
| 16 | L | S | L S | | | | | | | | |
| 17 | K | F | K F | | | | | | | | |
| 18 | Q | D | Q D | | | | | | | | |
| 19 | H | A | H A | | | | | | | | |
| 20 | H | L | H L | | | | | | | | |
| 21 | V | N | V N | | | | | | | | |
| 22 | E | F | E F | | | | | | | | |
| 23 | Y | V | Y V | | | | | | | | |
| 24 | G | T | G T | | | | | | | | |
| 25 | S | M | S M | | | | | | | | |
| 26 | D | L | D L | | | | | | | | |

(viii) The following formula was inserted into cell D1:

**=SUM(COUNTIFS($C1:C26, {"A C","A G","A R","A S","C A","C T","D I","D L","D V","E F","E L","F E","F K","G A","G P","G S","G T","H M","H V","I D","I N","I Y","L D","L E","L K","L N","L Q","K L","K F","M H","M Y","N I","N L","N V","P G","P R","P W","Q L","Q V","R A","R P","R S","R T","S A","S G","S R","S S","S T","T C","T G","T R","T S","T W","V D","V H","V N","V Q","V Y","W P","W T","Y I","Y M","Y V"})) .**

The C26 in the formula should be altered to match the last row containing amino acid pairs in column C.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | I | N | IN | 12 |
| 2 | I | D | ID | |
| 3 | A | V | AV | |
| 4 | G | V | GV | |
| 5 | K | F | KF | |
| 6 | N | V | NV | |
| 7 | S | I | SI | |
| 8 | G | D | GD | |
| 9 | V | S | VS | |
| 10 | D | G | DG | |
| 11 | E | K | EK | |
| 12 | A | V | AV | |
| 13 | F | K | FK | |
| 14 | F | E | FE | |
| 15 | V | K | VK | |
| 16 | L | S | LS | |
| 17 | K | F | KF | |
| 18 | Q | D | QD | |
| 19 | H | A | HA | |
| 20 | H | L | HL | |
| 21 | V | N | VN | |
| 22 | E | F | EF | |
| 23 | Y | V | YV | |
| 24 | G | T | GT | |
| 25 | S | M | SM | |
| 26 | D | L | DL | |

(ix) The final score calculated by this formula equals the amount of all the +1 amino acid pairs out of the total amino acid pairs within the alignment section between the SNC sequence and the subject protein. In this case 12/26 amino acid pairs were awarded a score of +1 which converts into an MR score of 46%.

(x) The amino acid sequence of Aß residues 32-7 corresponds to IIAGKNSGVDEAFFVLKQHHVEYGSD and the amino acid sequence of 3'-5' RNA helicase YTHDC2 isoform 1 residues 698-723 corresponds to NDVVFVIDSGKVKEKSFDALNFVTML. Aligning these in a table manually also allows scoring:

```
Aß      IIAGKNSGVDEAFFVLKQHHVEYGSD
YTHDC2 NDVVFVIDSGKVKEKSFDALNFVTML
Score  11001100000011001000111101
```

(xi) The same process was followed for all residue alignments, however depending on the source of the subject protein, AS35CN and AS53CN or AS35NC and AS53NC, the corresponding SCN or SNC peptide was used initially in steps (i) - (iii) detailed above on pages 33-34 respectively.

(7e) The following examples have alignments copied from txt files from BLAST searches (see section 6d, (iv) page 23) with SNC or SCN sequences inserted. Example (a) has been taken from an AS35NC BLAST search and Example (b) has been taken from an AS35CN BLAST search. The scoring has used table 7b above (page 31):

Example (a)

```
>pancreatic triacylglycerol lipase precursor [Homo sapiens]
Sequence ID: NP_000927.1 Length: 465
Range 1: 240 to 256

Score:25.8 bits(55), Expect:35,
Method:Compositional matrix adjust.,
Identities:10/17(59%), Positives:12/17(70%), Gaps:0/17(0%)

SNC     54   VELEKGVLPQLEQPYVF  70

Query   54   HLDLFPHDGVELVGIHK  70
             HLD FP+ GVE+ G  K
Sbjct   240  HLDFFPNGGVEMPGCKK  256
```

Scoring (a):

```
SNC     VELEKGVLPQLEQPYVF
Sbjct   HLDFFPNGGVEMPGCKK
Score   11111110111001001
```

Example (b)

```
>sterile alpha and TIR motif-containing protein 1 precursor [Homo sapiens]
Sequence ID: NP_055892.2 Length: 724
Range 1: 326 to 341

Score:23.5 bits(49), Expect:214,
Method:Compositional matrix adjust.,
Identities:11/16(69%), Positives:13/16(81%), Gaps:0/16(0%)

SCN     93   LEVVKHGHNTSLADSR  108

Query   93   DLQQFVPVLWSNRLRA  108
             DLQ+ VP+L SNRL A
Sbjct   326  DLQRLVPLLDSNRLEA  341
```

Scoring (b):

```
SCN     LEVVKHGHNTSLADSR
Sbjct   DLQRLVPLLDSNRLEA
Score   1110111010111101
```

The total MR score is best expressed as a % of the total number of residues in the target sequence, in this case %MR = 71% (12/17) for Example (a) and 75% (12/16) for Example (b). The higher this %MR the more likely a binding

interaction could potentially to occur. This number will always be equal to or greater than the % identity score from the BLAST search. Within the results table (see section 6h, page 27) a ninth column can be inserted with the heading % MR and the values determined from this scoring added.

| Human Protein that theoretically binds human Aß | Protein ID | Size | Residues | Aß residues | % ID | % +ve | % Gaps | % MR |
|---|---|---|---|---|---|---|---|---|
| 3'-5' RNA helicase YTHDC2 isoform 1 | NP_073739.3 | 1430 | 698-723 | 32-7 | 35 | 50 | 0 | 46 |
| 3'-5' RNA helicase YTHDC2 isoform 2 | NP_001332904.1 | 1268 | 536-561 | 32-7 | 35 | 50 | 0 | 46 |
| 3'-5' RNA helicase YTHDC2 isoform 3 | NP_001332905.1 | 1130 | 398-423 | 32-7 | 35 | 50 | 0 | 46 |

(7f) Where comparing multiple potential protein interactions the % MR is a useful indicator of potential interactions. The % scoring accounts for the size of the interacting region without consideration for the overall size of each protein.

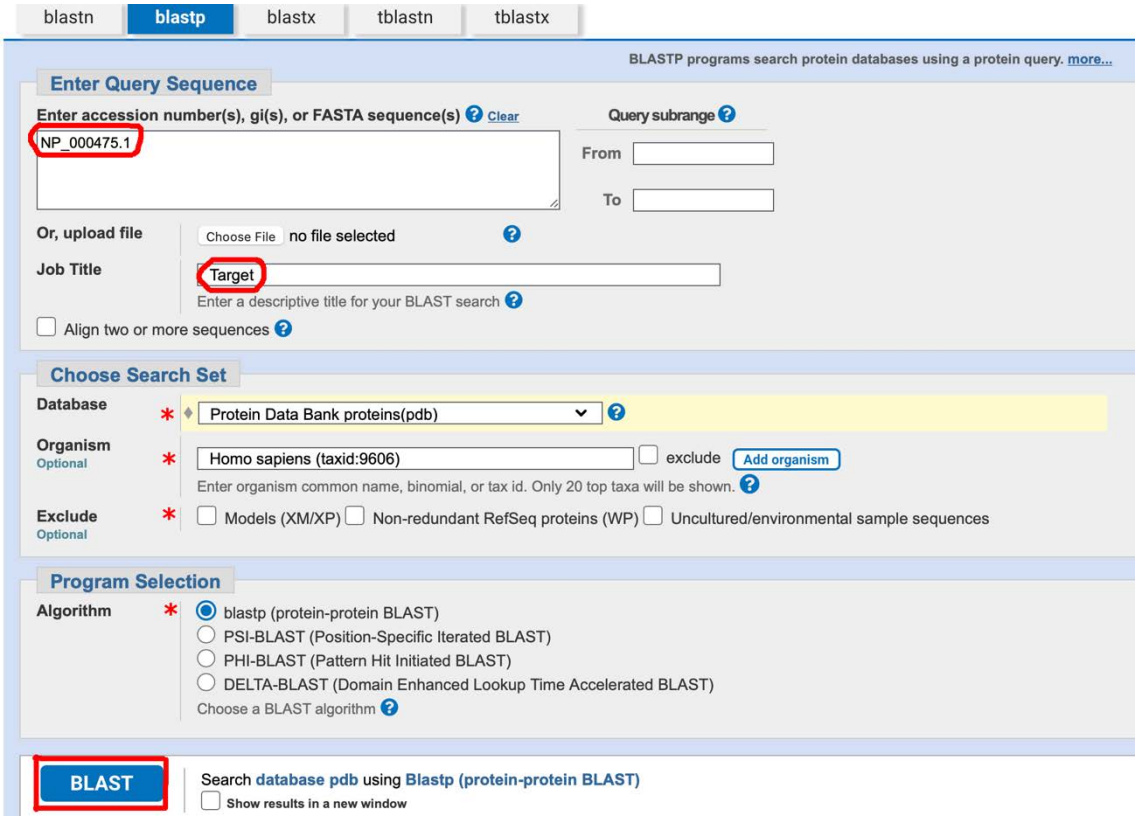
The size of each interacting region will be variable, and a larger interacting region may not be more important than a very short one. It is important to remember that protein interactions can involve both small and large sequences binding together. For example, the tripeptide thyrotropin-releasing hormone (TRH) binds to a 398 amino acid receptor but only interacts with a very limited number of the TRH receptor amino acids. As such it is important to consider short interacting regions as well as long interacting regions.

The nature of the molecular recognition theory does not directly take into account secondary or tertiary structures of the proteins that might interact and the %MR should always be considered as a predictive tool with these limitations in mind.

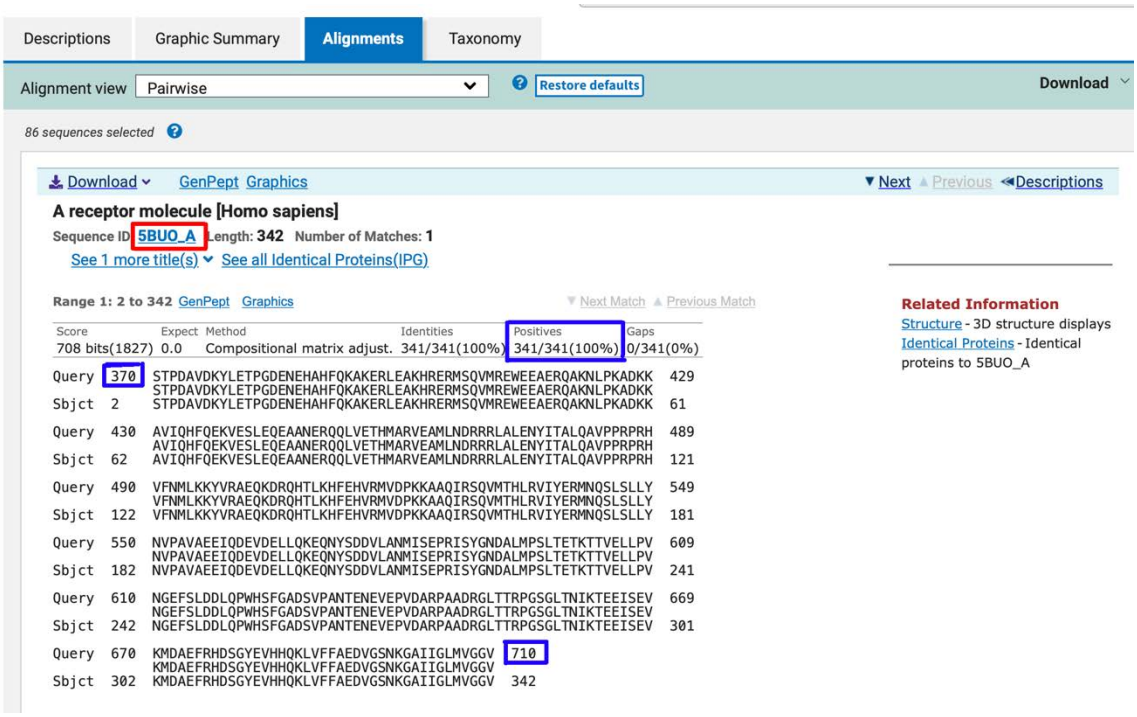
# 8: PDB files for protein-protein interaction modelling

(8a) For molecular modelling using the ZDock (http://zdock.umassmed.edu, section 9 pages 52-55 below) the software requires a PDB file from the target protein and a PDB file from the suggested interacting protein. From the table of results from the Molecular Recognition analysis above (section 7e, page 39) it is possible to select results in terms of the regions of proteins that interact. Choosing which interactions to study is detailed in Section 12a-e (pages 73-74).

(8b) Since the target protein will have a protein ID (NP_*** number: see section 2b, page 7 above) and the interacting proteins from the BLAST search results will also have a protein ID (NP_*** number: see results tables generated in section 7e, page 39 above) it is possible to run a Protein BLAST search to identify similar protein structures. Under the header "Enter accession number, gi or FASTA sequence" paste in the protein ID for the target or interacting protein, in this example the NP_000475.1 protein ID has been used (circled in red). Also enter the protein name into the Job title box (circled in red). Under the "Choose Search Set", "Database" select Protein Data Bank proteins(pdb), under the "Organism" type in homo sapiens and select "Homo sapiens (taxid:9606)", leave the tick boxes for Exclude "Models (XM/XP)", "Non-redundant RefSeq proteins (WP)" and "Uncultured/environmental sample sequences" unchecked plus select Algorithm "blastp (protein-protein BLAST)" – see red * marks below:

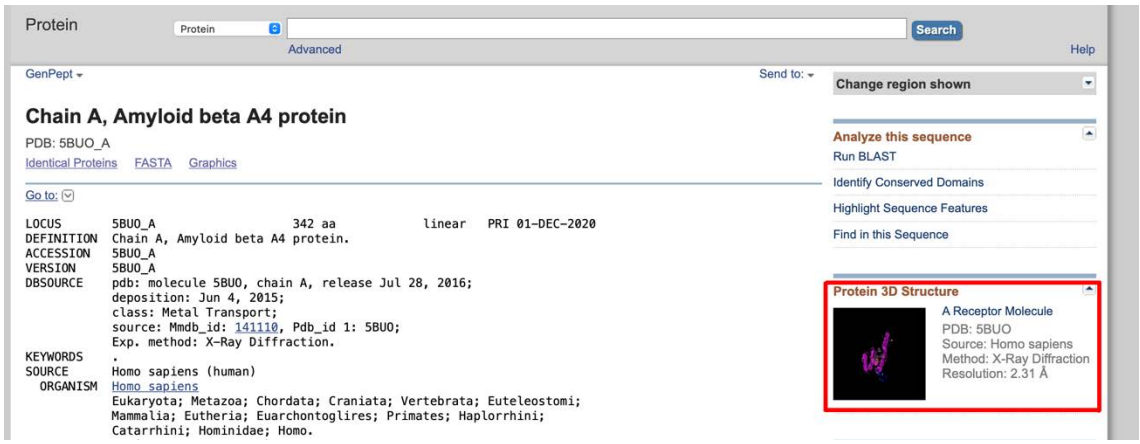


After clicking BLAST (circled in red above) a series of results will appear. Using the "Alignments" tab in the results it is possible to find structures similar to regions of interest:



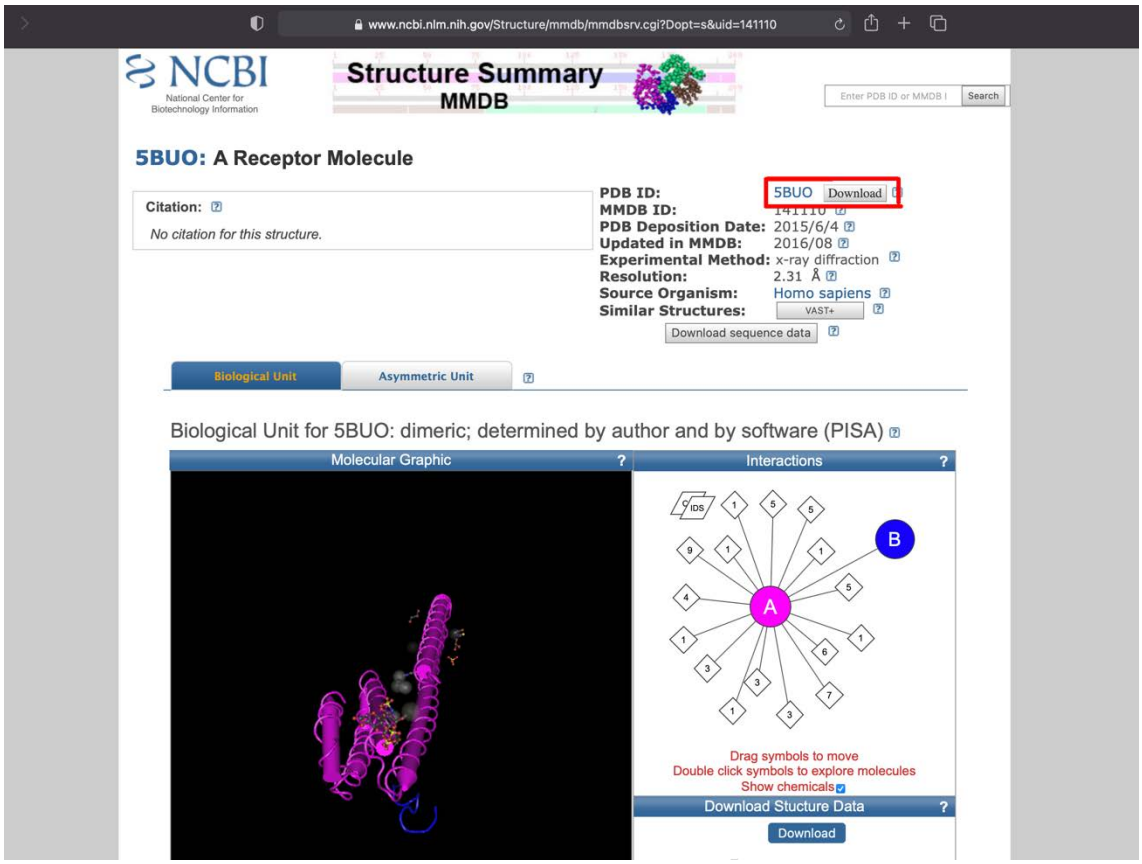

Checking through the alignments allows identification of the regions of the searched protein that are present within a structure and therefore choice of a structure that will cover the region of interest from the antisense peptide BLAST results. The protein encoded by NP_000475.1 is the 770 amino acid amyloid precursor protein. From the alignment's information above the A receptor molecule with a sequence ID 5BUO_A (circled in red) has a region similar to the NP_000475.1. The results show that the NP_000475.1 residues 370 – 710 (circled in blue) are 100% identical (Positives circled in blue) to the 5BUO_A residues 2-342.

Clicking on the 5BUO_A link (circled in red) will go to information about the protein, which also contains links to the PDB files (circled in red below):



For many of the protein entries similar links to structures are available in the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/), often with links to multiple structures. The BLAST search detailed above is a simple way of selecting an appropriate structure that covers a region of interest.

Clicking on the link for the Protein 3D Structure (circled in red above) will go to the structure file. Within this file there is an option to download the PDB file (see link circled in red below) which can then be saved for use in protein modelling as detailed in sections 9 pages 52-55 below:




(8c) The ZDock molecular modelling software uses PDB files from an individual target protein and suggested interacting protein. The key pieces of information from the antisense peptide BLAST searches required to aid the molecular modelling are the residue number information contained within the results tables generated in section 7e (page 39). The residues of the target protein (in this example amyloid-ß residues 32-7) involved in binding and the residues of the interacting protein (in this example NP_073739.3 - 3'-5' RNA helicase YTHDC2 isoform 1 residues 698-723) from the BLAST search results:

| Human Protein that theoretically binds human Aß | Protein ID | Size | Residues | Aß residues | % ID | % +ve | % Gaps | % MR |
|---|---|---|---|---|---|---|---|---|
| 3'-5' RNA helicase YTHDC2 isoform 1 | NP_073739.3 | 1430 | 698-723 | 32-7 | 35 | 50 | 0 | 46 |
| 3'-5' RNA helicase YTHDC2 isoform 2 | NP_001332904.1 | 1268 | 536-561 | 32-7 | 35 | 50 | 0 | 46 |
| 3'-5' RNA helicase YTHDC2 isoform 3 | NP_001332905.1 | 1130 | 398-423 | 32-7 | 35 | 50 | 0 | 46 |

A BLAST search as described in section 8b (pages 40-42 above) for the NP_073739.3 sequence corresponding to isoform 1 of the 3'-5' RNA helicase YTHDC2 sequence which interacts with Aß (see table in 8c, page 42 above) identified two structures derived from the YTHDC2 sequence. These structures 6K6U (https://www.ncbi.nlm.nih.gov/Structure/pdb/6K6U) and 2YU6 (https://www.ncbi.nlm.nih.gov/Structure/pdb/2YU6) which both cover a structure representing the YTH domain which corresponds to residues 1288-1421 of isoform 1 of the 3'-5' RNA helicase YTHDC2 sequence. This region of 3'-5' RNA helicase YTHDC2 is outside the proposed 698-723 region of interaction with Aß and therefore modelling using these models is not possible. There are PDB structures for other proteins which do show similarity to the appropriate region of YTHDC2 (698-723), but they are not identical proteins.

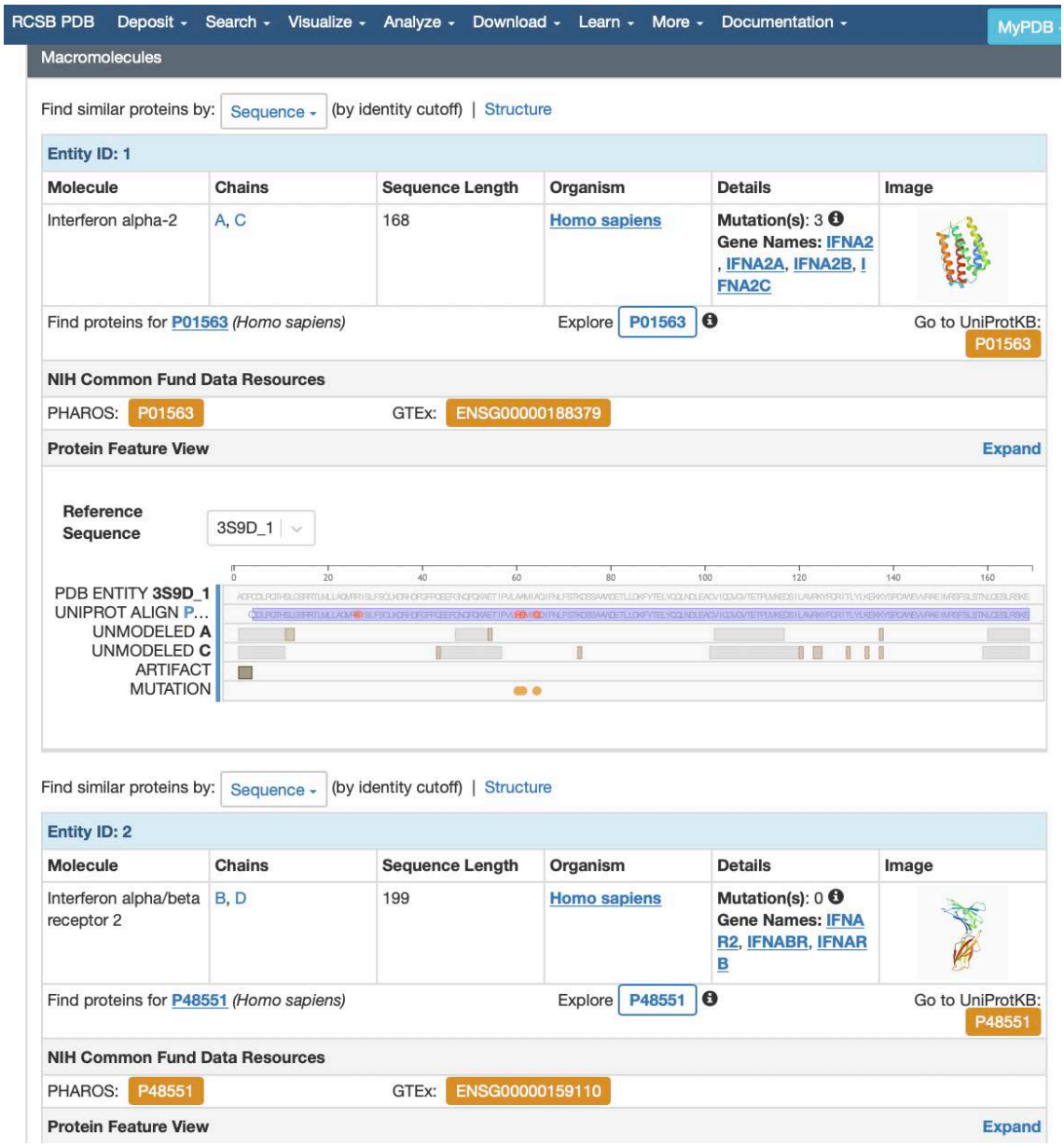
(8d) For protein sequences, like the 3'-5' RNA helicase YTHDC2 698-723 region, where no PDB file is available a predicted protein structure can be created, using the I-Tasser website (https://zhanglab.ccmb.med.umich.edu/I-TASSER/). Protein sequences of 10-1500 amino acids can be entered and have a predicted structure created. There are also options to specify related known structure files as templates if these are available, with details in the dropdown Option menus. The site requires registration as a user, using an academic email address, and download links to the created files will be sent via email.

(8e) PDB file information can be obtained from a number of sources, the RSCB Protein Databank (https://www.rcsb.org) is the preferred choice and contains validated structural information. Structures can also be obtained from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/), the UniProt protein database (https://www.uniprot.org), or the Model Archive (https://www.modelarchive.org).

(8f) Where structures for a protein are only available in the crystallographic information file (CIF) or macromolecular CIF (mmCIF) formats, which download as "name.cif" files these can be converted to PDB format using the PDBx/mmCIF conversion service (https://mmcif.pdbj.org/converter/index.php?l=en).

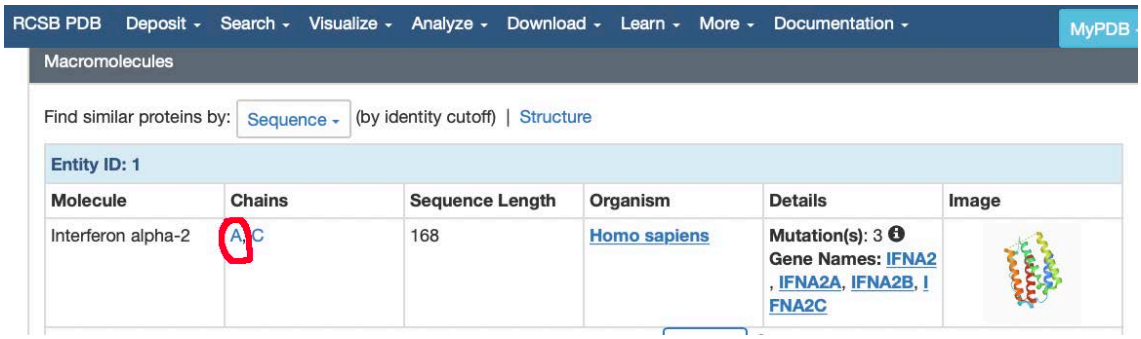
(8g) For many structural models the PDB files contain multiple protein chains from one or more proteins. The PDB file for human catalase (PDB 1DGH; https://www.rcsb.org/structure/1DGH) has four catalase chains as the biologically active enzyme exists as a tetramer of the molecule (Putnam *et al.,* 2000).

An example of a structure with more than one protein type is the structure of the human interferon alpha-2 in complex with human interferon alpha/beta receptor 2 PDB file 3S9D, http://www.rcsb.org/structure/3s9d). The structure comprises two human interferon alpha-2 chains (A & C) in complex with two human interferon alpha/beta receptor 2 chains (B & D):




(8h) An important feature of PDB files that must be checked is the numbering of residues. PDB files are derived from structural information obtained using purified proteins, which may represent fragments of the whole molecule or post-translationally modified proteins and may therefore lack regions cleaved as part of this process. Details of how PDB files are derived is available from https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure. Often the PDB residue numbering differs from the NCBI or Uniprot residue numbering. Structures are often analysed after removal of signal sequences and some have other regions modified. In the following example the 3S9D_A chain described above will be used. It is important to check the sequence numbering of the PDB chain against the NCBI chain identified in BLAST searches described above (see section 6 pages 20-30 and section 7 pages 31-39 above). In this example the A chain information can be found at http://www.rcsb.org/structure/3s9d.

Selecting the Interferon alpha-2 A chain (circled in red) will access information about that specific chain:

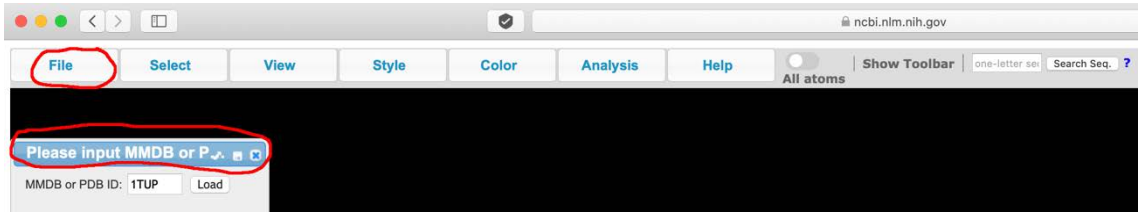


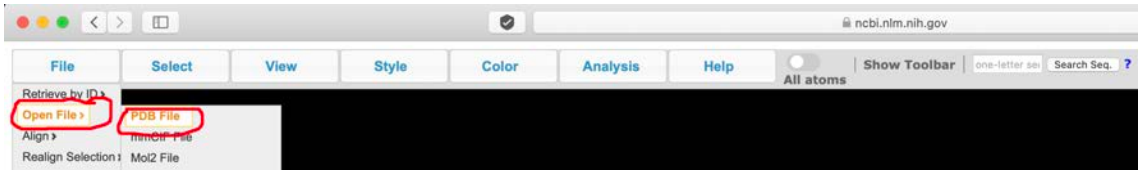

In this example the sequence comparison is between the Interferon alpha-2 A chain and the Uniprot P01563 file (https://www.uniprot.org/uniprot/P01563), circled in red. This sequence corresponds to the NCBI P01563 file (https://www.ncbi.nlm.nih.gov/protein/P01563/). Selecting the K residue highlighted with a red dot (see blue arrow on diagram) on the Uniprot P01563 line of the information brings up numbering information for that specific residue (circled in blue) and compares the Uniprot numbering with the PDB model numbering, selecting any residue in the sequence would have the same effect:

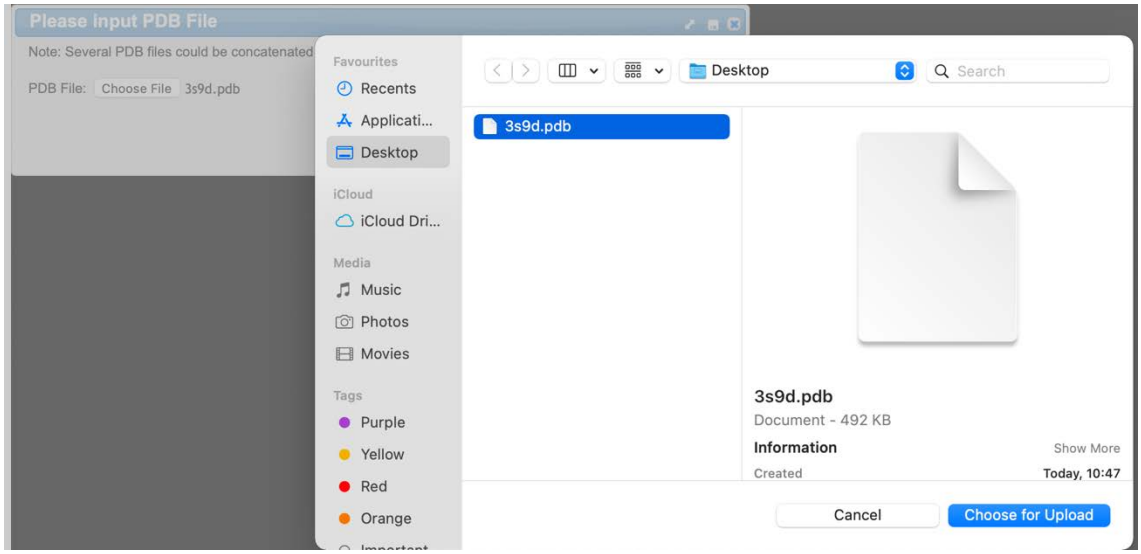

The numbering in the blue circled box shows that residue 26 of the model corresponds to the Uniprot P01563 number of 46 (the author of the model used a slightly different numbering scheme starting at -2, which results in an author residue number of 23 for this residue and this numbering shows when using structure viewers (see section 8i, page 46 below). As such if a BLAST search had suggested that residues 50-65 of the Interferon alpha-2 interacted with another protein when selecting these residues on this model they would correspond to residues 30-45 (or author residues 27-42). It often easier to select residues based on a sequence rather than numbers. In this example the sequence would correspond to FSCLKDRHDFGFPQEE. Such sequence information can be found in the results downloaded from BLAST searches if they show similarity to either the AS35CN, AS35NC, AS53CN or AS53NC sequences (see section 6d (iii) Alignments and (iv) Text file on pages 22-23 above, labelled as Sbjct) or they can be found in the SNC sequence of the target protein using the protein id link detailed in section 2b, page 7 above).

(8i) Within the ZDock modelling software (see section 9 pages 52-55 below) it is not possible to select different chains from uploaded files and therefore all protein chains within a model will be modelled. As part of the modelling, it may be preferable to create a PDB file with only a single chain within it (or even part of a chain). The PDB files can be viewed and edited using the iCn3D protein structure viewer (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html) as described by Wang *et al.,* (2020). See Help Doc (circled in red) and Selection Hints (circled in blue).


(8j) Initially close the "Please input MMDB or PDB" window if it is open and go to File on the main window. From the File menu select Open file from the dropdown and then PDB from the second dropdown to upload a saved PDB file:

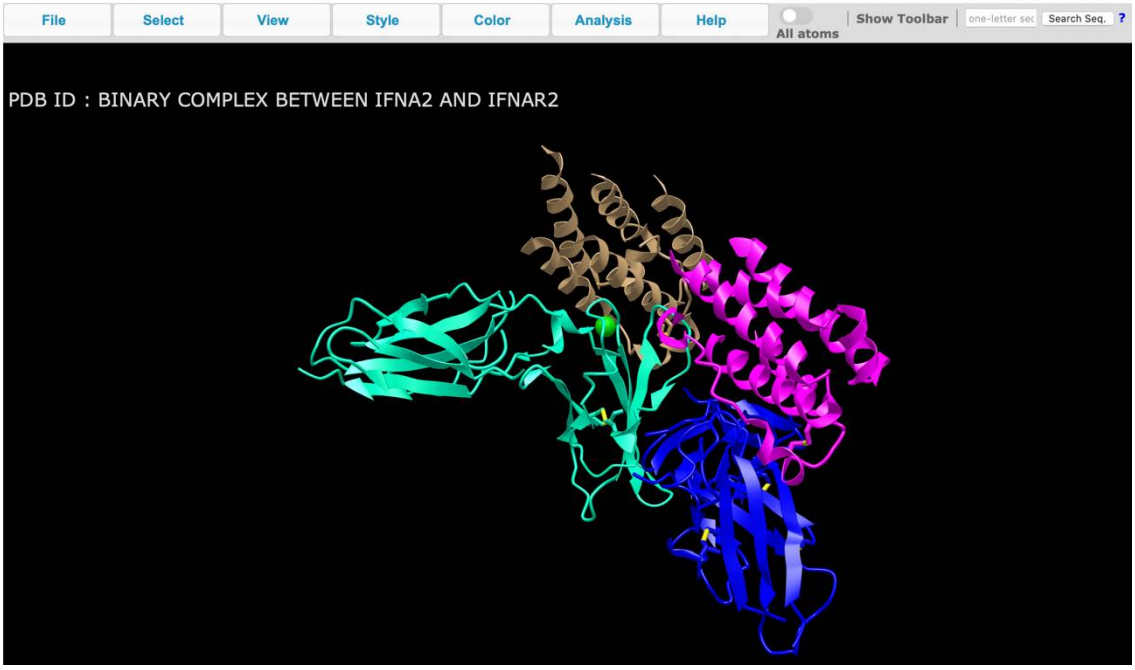


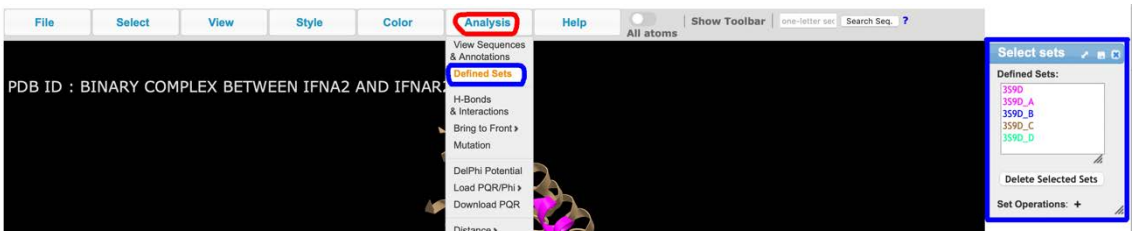



(8k) Having chosen the PDB file (3s9d.pdb in this example) click on the load (circled in red):


The initial view of the uploaded structure will look as follows:




Selecting Analysis (circled in red) and "Defined Sets" (circled in blue) from the dropdown menu will result in the appearance of a box listing the chains of the PDB file (outlined in blue on the RHS).



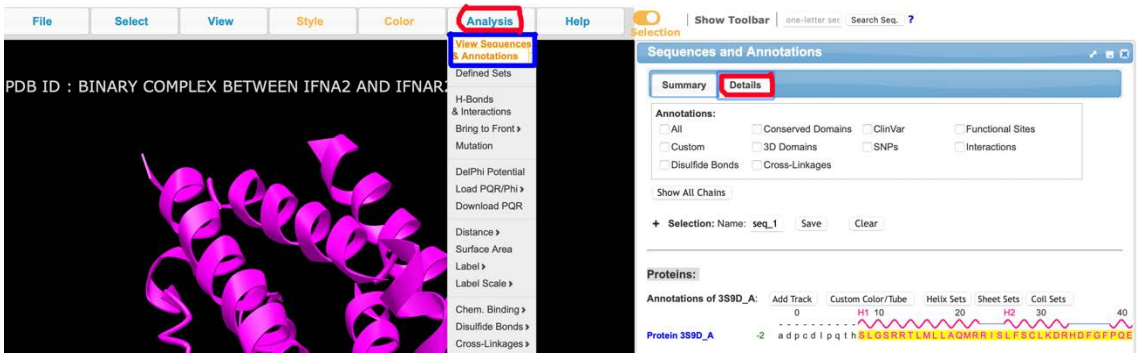

The Defined Sets correspond to the different chains of the structure. In this example the 3S9D_A and 3S9D_C correspond to the Interferon alpha-2 A and C chains, whilst the 3S9D_B and 3S9D_D correspond to the Interferon alpha/beta receptor 2 B and D chains (see section 8g; page 44 above).

For this example, the 3S9D_A chain will be selected (corresponding to the one of the Interferon alpha-2 molecules, represented by the A chain in the 3S9D model - http://www.rcsb.org/structure/3s9d). By clicking on the selected chain

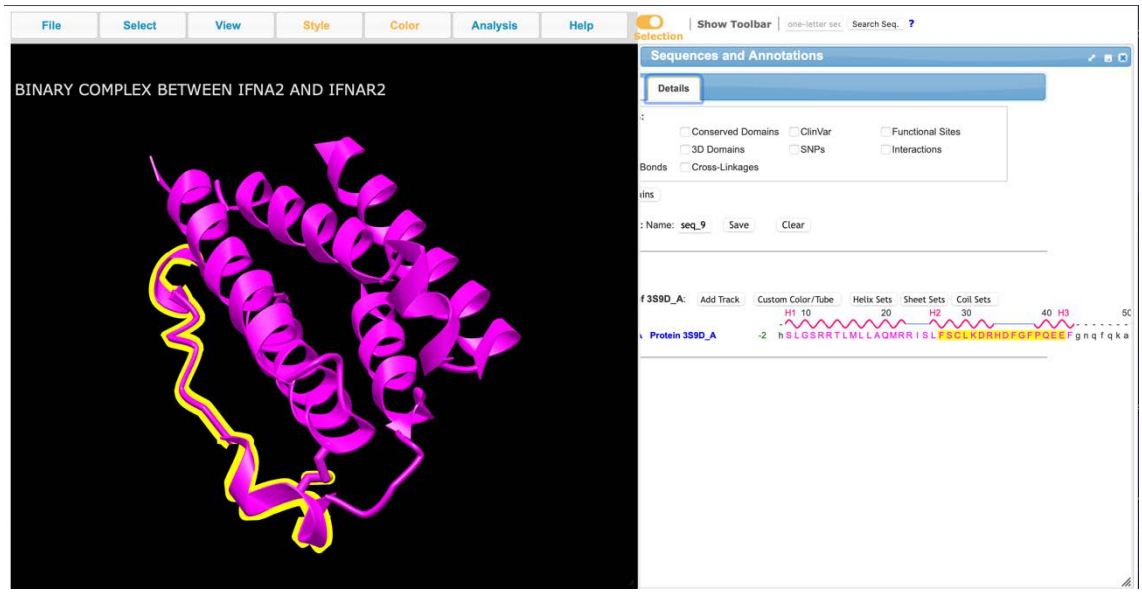
(circled in red) followed by the View menu (circled in red) and then the "View only selection" option in the dropdown menu (circled in blue) only the A chain, will be visible – the pink chain in this model:


Files uploaded into the iCn3D protein structure viewer (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html) can be further modified by selecting specific parts of chains to create a representation of a specific region. If this option is not required proceed straight to section 8m on page 49 below.

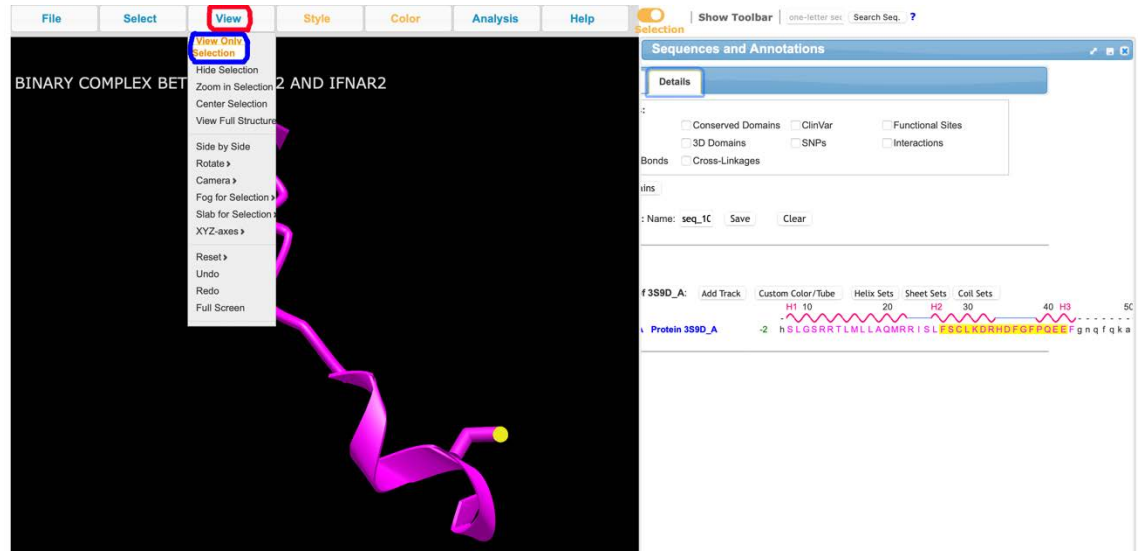
(8l) After closing the "Select Sets" window selecting the Analysis (circled in red) followed by "View Sequences & Annotations" from the dropdown menu (circled in blue) will bring up a box showing the chain information. Selecting the "Details" tab in this box (circled in red) will bring up the sequence details:




The protein sequence for the model is highlighted in Yellow at the bottom of the "Details" window. Scrolling across this allows the area of interest to be found, in this case the "FSCLKDRHDFGFPQEE" sequence detailed in section 8h above on page 45. Clicking on the F at the start of this sequence and dragging across to the E at the end allows that specific sequence to be selected and the residues will also be highlighted in yellow on the pink image to the left:




Selecting "View" (circled in red) in the main window followed by "View Selection Only" (circled in blue) in the dropdown menu will result in an image of just the selected structure components:



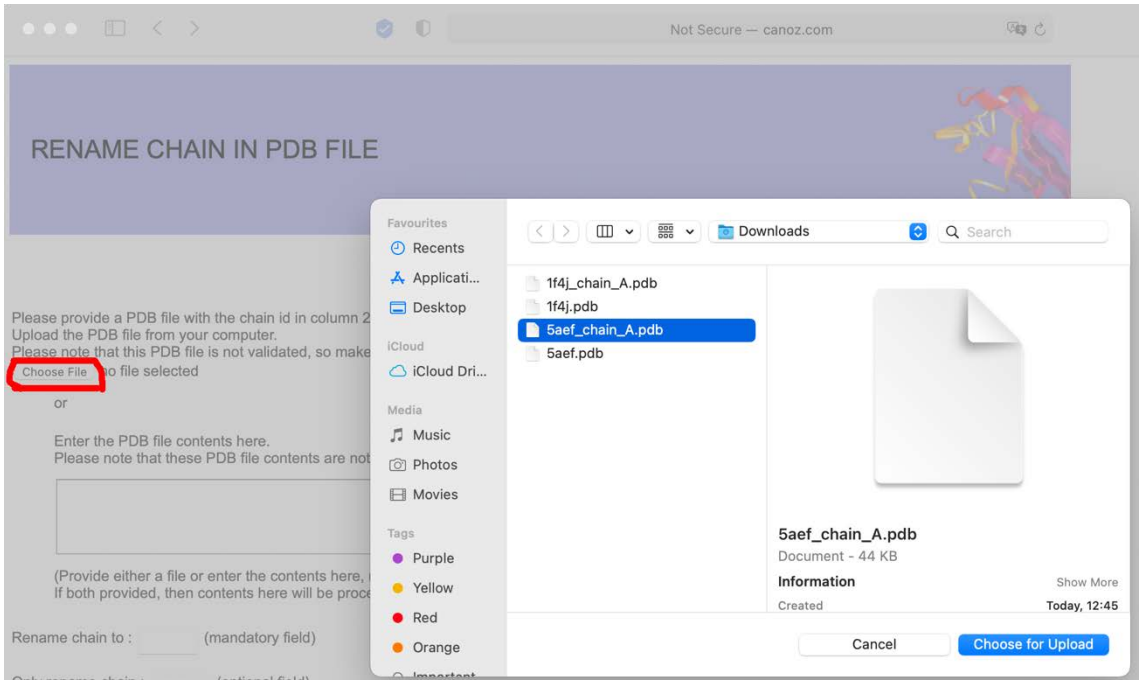

(8m) Clicking on the File (circled in red) followed by "Save Files" (circled in blue) in the first dropdown menu, followed by PDB (circled in blue) in the second dropdown menu will result in the download of a new PDB with just the selected chain or parts of a chain information within the file:

The downloaded file can be renamed from "custom_icn3d.pdb" to an appropriate name, for example 3S9D_A_chain or 3S9D_A_binding_site. The saved PDB file can be used for modelling interactions in the ZDock program as described in section 9 (pages 52-55 below).

(8n) The ZDock program uses two PDB files and has a specific requirement that the chains of each protein within the PDB file have different labels. If there are two A chains these will be merged in the resultant comparison model and this will affect the analysis and production of images.

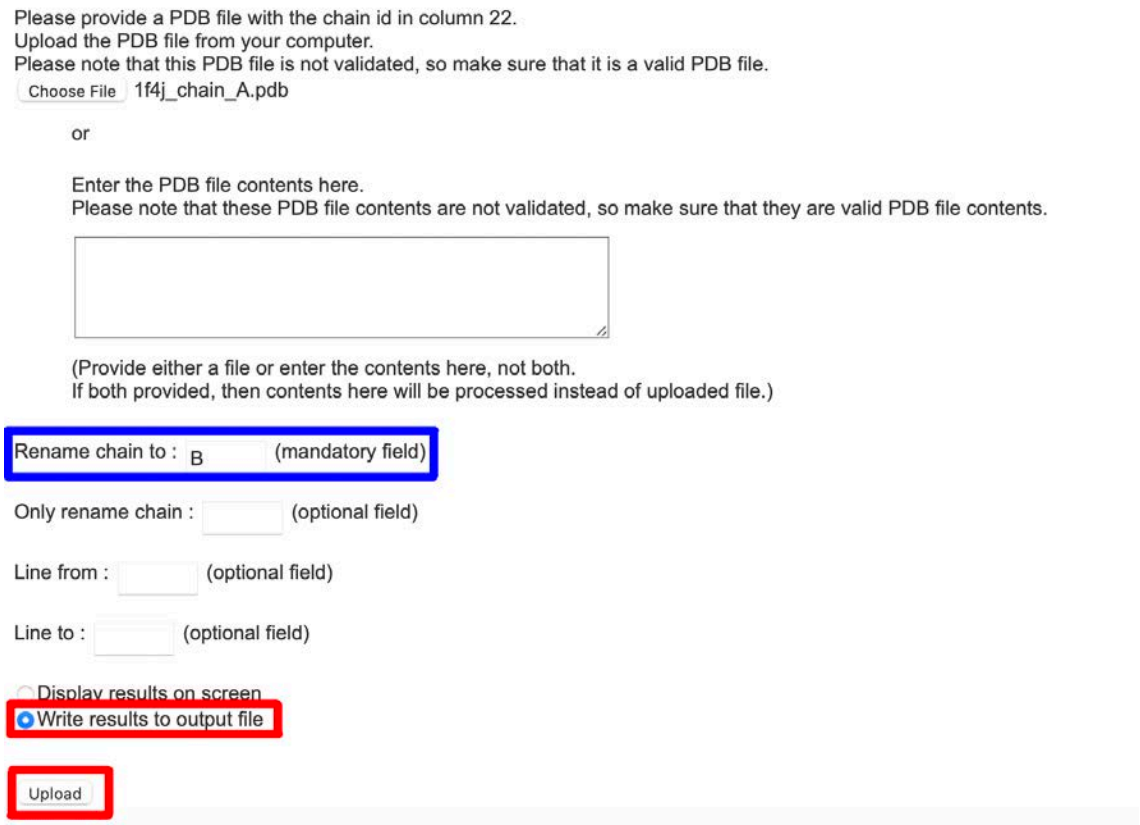
For an example, the catalase PDB 1F4J (https://www.rcsb.org/structure/1F4J) and the amyloid-ß PDB 5AEF (https://www.rcsb.org/structure/5AEF) could be used as source PDB files, based on the interaction between human catalase and amyloid-ß (Milton *et al.,* 2001). The catalase PDB 1F4J file represents a tetramer of 4 chains labelled A, B, C and D. The amyloid-ß PDB 5AEF file represents a dimer of 2 chains labelled A and B. If when the PDB files were prepared as detailed above resulting in one PDB containing the catalase A chain (1F4J_chain_A) and one PDB containing the amyloid-ß A chain (5AEF_chain_A) this would create problems in the ZDock modelling (see section 9 pages 52-55 below).

To overcome these problems of PDB files containing the same chain labels the renaming chains website (http://www.canoz.com/sdh/renamepdbchain.pl) can be used. After opening the site files can be directly loaded (circled in red), for this example the 1F4J_chain_A is used to rename the chain A:




In the "Rename chain to" option box the new chain label name is entered, in this case B (circled in blue), the "Write results to output file" option (circled in

red) is selected, and the "Upload" button (circled in red) then clicked to download a file called "renamepdbchain.pl":

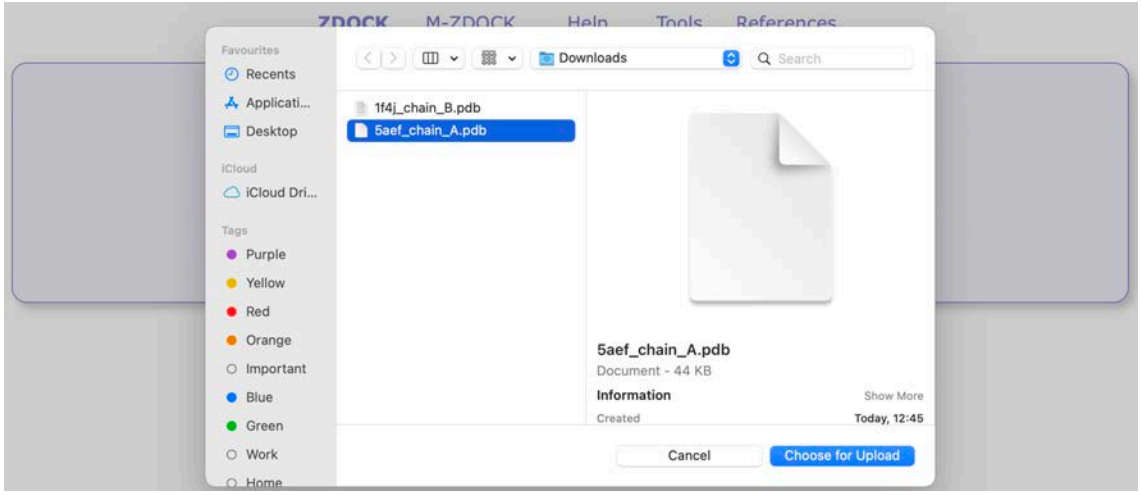



The "renamepdbchain.pl" file can be renamed, for example 1F4J_chain_B, and then be used in ZDock analysis (see section 9 pages 52-55 below).

# 9: ZDock protein-protein interaction modelling

(9a) The ZDock protein docking program (http://zdock.umassmed.edu), see Pierce *et al.,* (2011 & 2014), uses PDB structure files from two proteins to determine a model structure for two proteins binding each other. Launching the ZDock protein docking program (http://zdock.umassmed.edu) goes to a window where the PDB Files can be uploaded. The ZDock system requires an academic or registered non-profit email address for non-commercial access. The email address should be entered and a link to the files generated by ZDock will be sent to this address. Leave the select ZDock version as the default that is showing (in this case ZDock 3.02 – which is the latest version). The dropdown menu (circled in red) next to the Input Protein 1 or 2 is used to select PDB files that can be uploaded (make sure the PDB File is ticked and highlighted blue. The Choose File tab (circled in blue) should then be clicked:

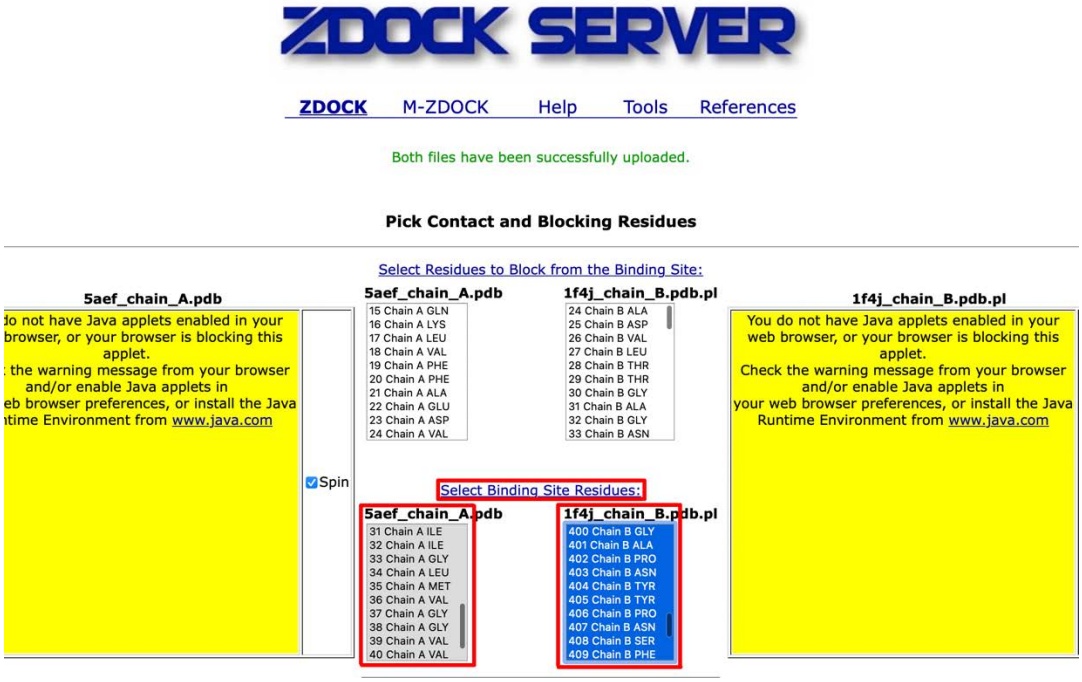

For this example, the catalase PDB file for 1F4J_chain_B and the amyloid-ß PDB 5AEF_chain_A have been uploaded:




Once both files have been uploaded click Submit (circled in red):


The next stage is to select the residues that are to be part of the binding site. These are the residues from the BLAST search results file plus the additional sequence information see Table in section 7e above (page 39; how these results can be obtained is fully detailed in sections 6 and 7 pages 20-39 above).

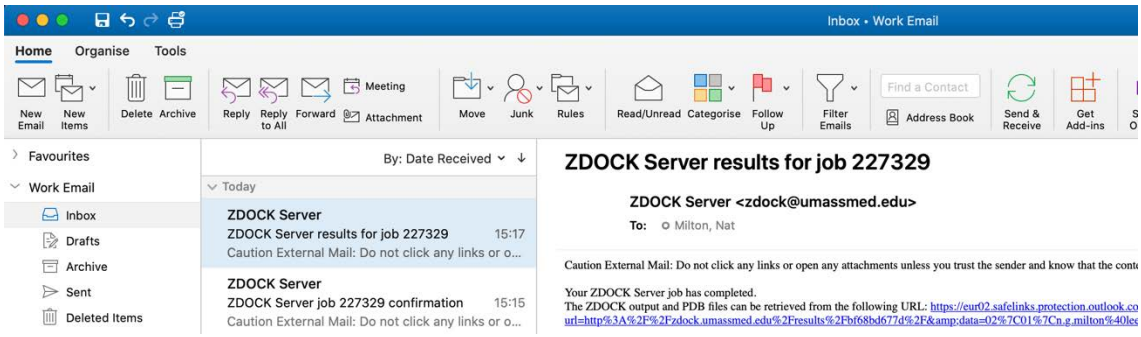
In an example of amyloid-ß and catalase the residues selected for the ZDock predictions are taken from Figure 1a of Milton *et al.,* 2001. The residues thought to interact based on a BLAST search are amyloid-ß 31-40 (circled in red for 5AEF_chain_A) and human erythrocyte catalase 400-409 (circled in red for 1F4J_chain_B). These residues are selected under "Select Binding Site Residues" on the ZDock server:



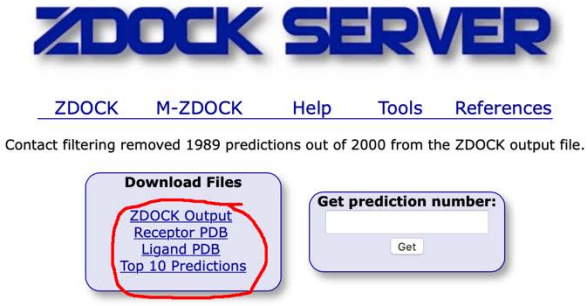

After clicking the "Submit" button at the bottom of the page the option to check the selected residues before final submission to ZDock is available.


After final submission to ZDock a notification will appear detailing the sending of results via email and the average wait time for receipt of the results.


A receipt email from ZDock Server, followed by second email with a link to the results will then be sent to the address provided (check junk email folder if not received). The job number (in this example 227329 should be noted and linked to the submitted PDB file information for future review):



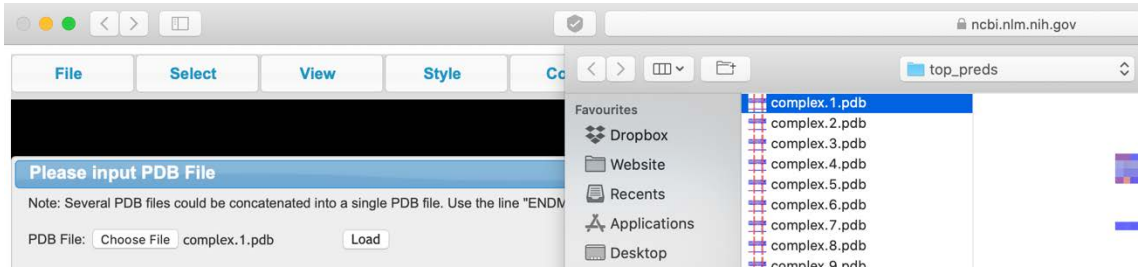

(9b) The link in the email is for the website results where the ZDock Output, Receptor PDB, Ligand PDB and Top 10 Predictions files should all be downloaded. In this example the Receptor PDB corresponds to amyloid-ß (5AEF chain A) and the Ligand PDB corresponds to catalase (1F4J chain B). The predictions are for the interaction between these two proteins.



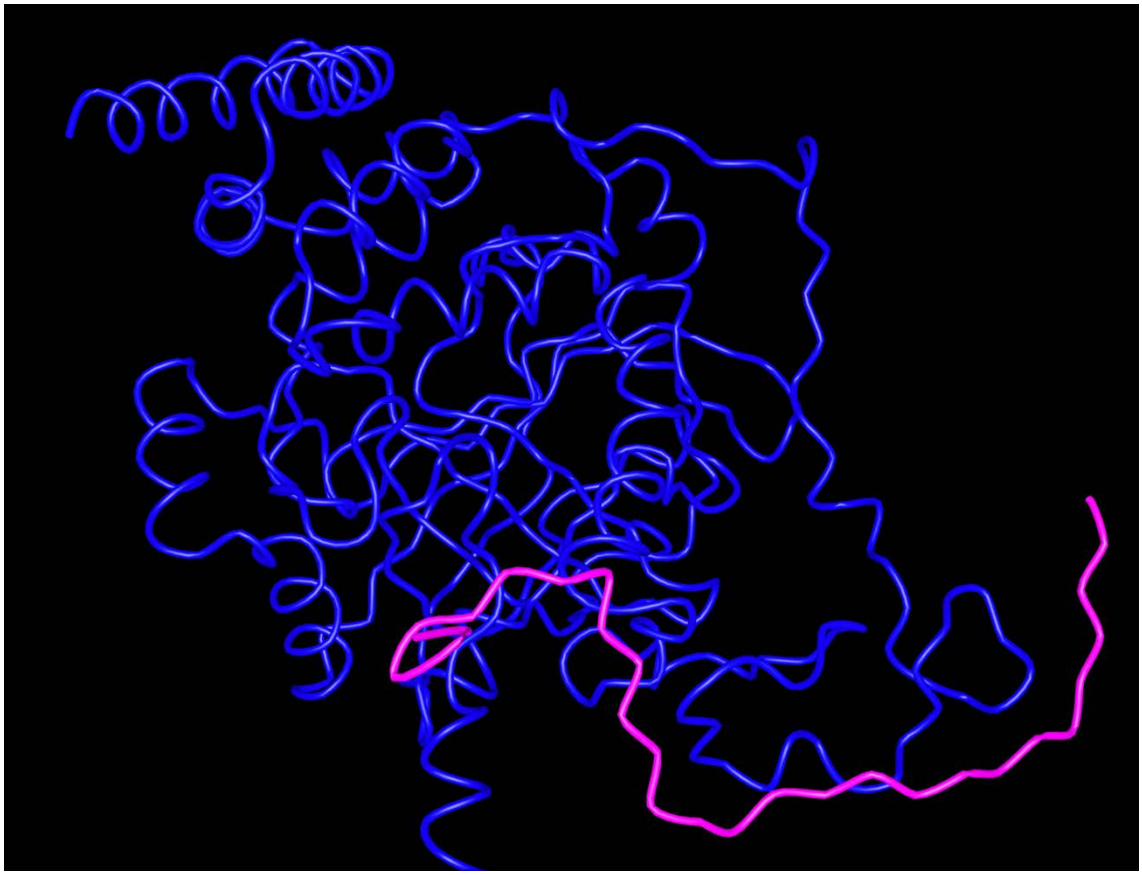

The top 10 predictions file will be downloaded as a Zip file which should be extracted and will give a folder labelled top_preds containing up to 10 files labelled complex1.pdb, complex2.pdb etc, which should be saved along with the other downloaded files in a suitable folder.

# 10: iCn3D protein-protein interaction data extraction

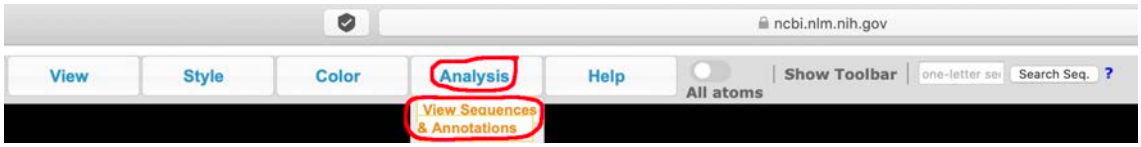
(10a) The top 10 prediction PDB files from the ZDock analysis (see section 9b; page 55 above) can also be viewed and modified using the iCn3D protein structure viewer (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html) as described in section 8i (above; page 46). The complex1.pdb file should be selected at first from the folder containing the downloaded ZDock files. This complex1.pdb contains the predicted complex between the two proteins with the highest ZDock score (see Help and References on the ZDock site, http://zdock.umassmed.edu).



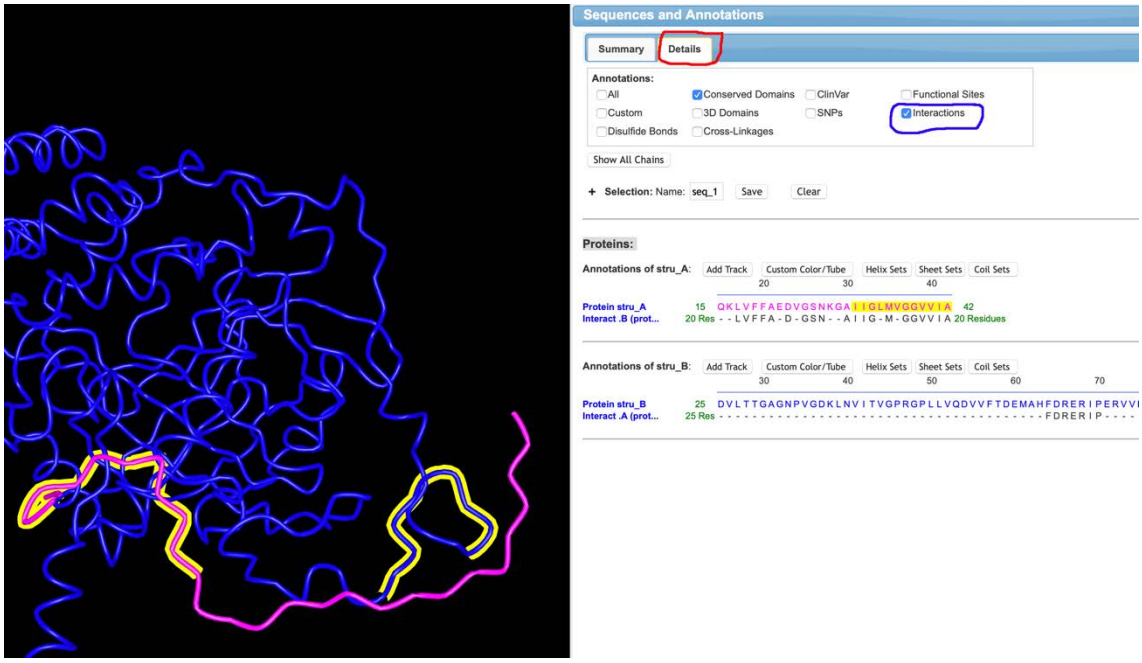

After selecting load a structure image should appear (these can take time to load).



(10b) Selecting the Analysis tab followed by the View sequences and annotations in the dropdown allows viewing of some of the information about the displayed structure.




Within the Sequences and Annotations window that appears selecting the details tab (circled in red) allows the residues of interest to be highlighted by clicking and dragging over the region from the first to the last residue of interest – in this example derived from Figure 1a of Milton *et al.,* 2001 the IIGLMGGVV sequence that corresponds to the 31-40 residues of amyloid-ß in Protein stru_A has been selected which is then highlighted in yellow. The region selected also shows as a yellow selection on the model. By doing the same for each protein, in this case selecting the GAPNYYPNSF sequence that corresponds to the residues 400-409 of catalase in Protein stru_B, it is possible to view how close the BLAST search sequence interactions from the results (for example see Table in 9a page 42 above). Selecting the Interactions box (circled in blue) shows the residues within the displayed model that interact.

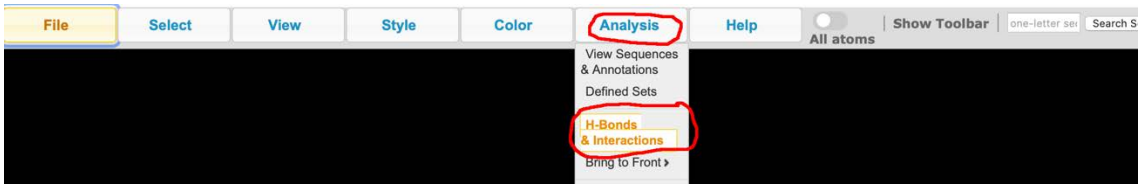



For comparison the image for complex3.pdb shows much closer alignment of the amyloid-ß 31-40 and catalase 400-409 highlighted sequences but also shows considerably fewer residues interacting (see green text in the Interact.B and Interact.A lines in the image above (complex1.pdb) and image below (complex3.pdb). In these examples for amyloid-ß (11 residues for complex3.pdb and 20 residues for complex1.pdb) and catalase (8 residues for complex3.pdb and 25 residues for complex1.pdb). The lower the complex number the more valid the model is in terms of the ZDock algorithm scoring, even if this may not

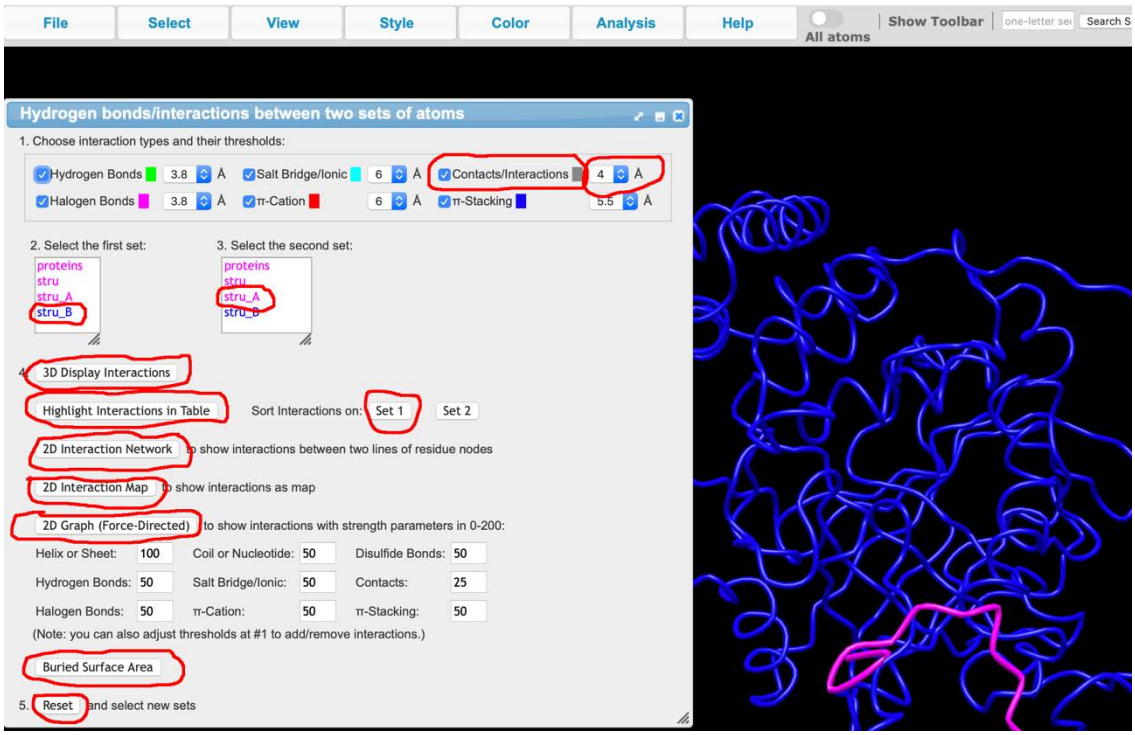
be as close to the predicted alignment from the antisense similarities used to suggest the interaction.


(10c) By selecting the download feature in the Sequences and Annotations window (circled in red) an html file containing the interactions information is downloaded and can be saved to the folder with the downloaded ZDock files.


(10d) Selecting the Analysis tab followed by the H-Bonds & Interactions in the dropdown allows viewing of some of the information about the interactions between the two proteins in the displayed structure.




This will bring up a new window entitled "Hydrogen Bonds/Interactions between two sets of atoms:

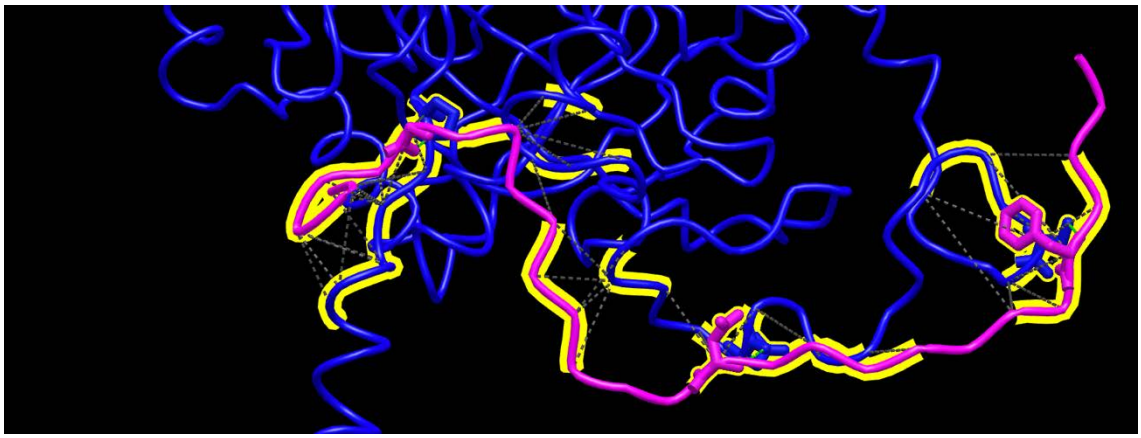



Within this window all the options are selected by default in part 1 "Choose interaction types and their thresholds". If for example only the Contacts/Interactions option had been ticked in the following parts the data for the other interactions would not be included.

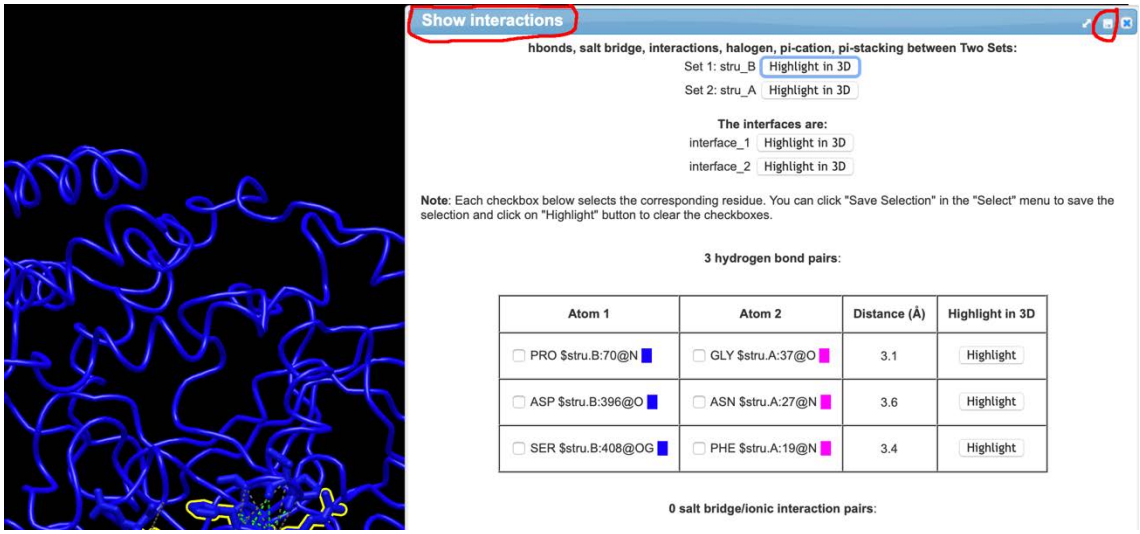
In part 2 the option to select the first set is given, for this example stru_B could be selected, which corresponds to the part of the model showing as blue lines and represents the catalase structure derived from catalase 1F4J PDB file used in the ZDock comparison (see section 9b, page 55 above).

In part 3 the option to select the second set is given, for this example stru_A could be selected, which corresponds to the part of the model showing as pink lines and represents the amyloid-ß structure derived from amyloid-ß 5AEF PDB file used in the ZDock comparison (see section 9b, page 55 above).

Part 4 allows the interactions to be views in a number of ways. The "3D Display Interactions" highlights the residues from catalase (blue) and amyloid-ß (pink) that interact and shows lines between the interacting amino acids:

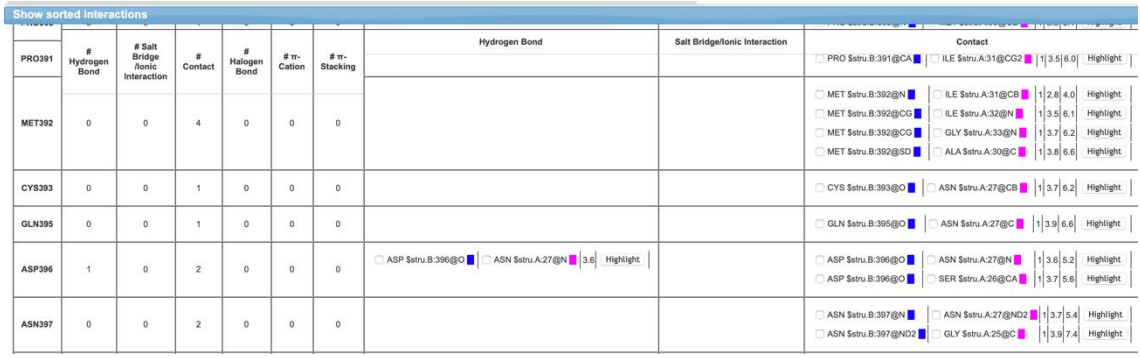


10(e) The "Highlight Interactions in Table" will generate a table of the interactions in a separate window:

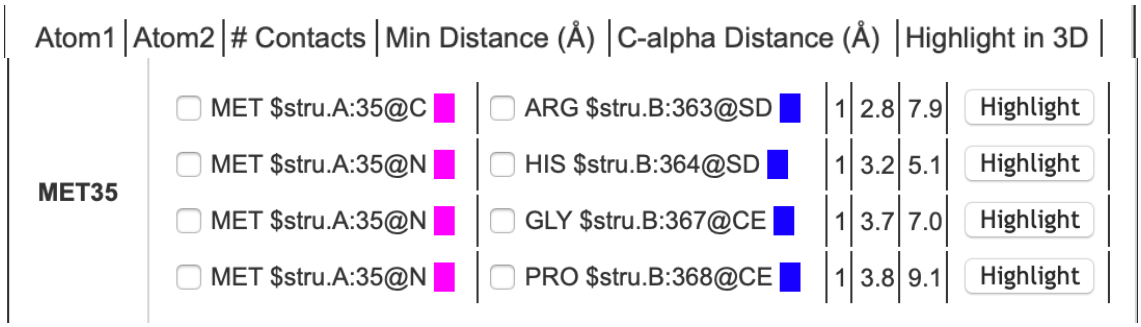



By selecting the download feature in the Show Interactions window (circled in red) an html file containing the interactions information is downloaded and can be saved to the folder with the downloaded ZDock files.

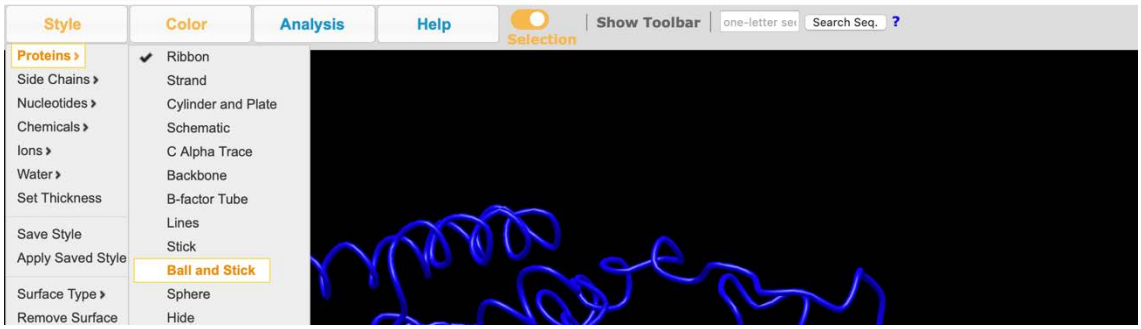
If the "Sort Interactions on: Set 1" is selected this table will be sorted based on the stru_B numbering, which corresponds to the part of the model showing as blue lines and represents the catalase structure derived from catalase 1F4J PDB, rather than sorted into separate tables for each interaction type. If "Sort Interactions on: Set 2" is selected the table will be sorted based on the stru_A numbering, which corresponds to the part of the model showing as pink lines and represents the amyloid-ß structure derived from 5AEF PDB.



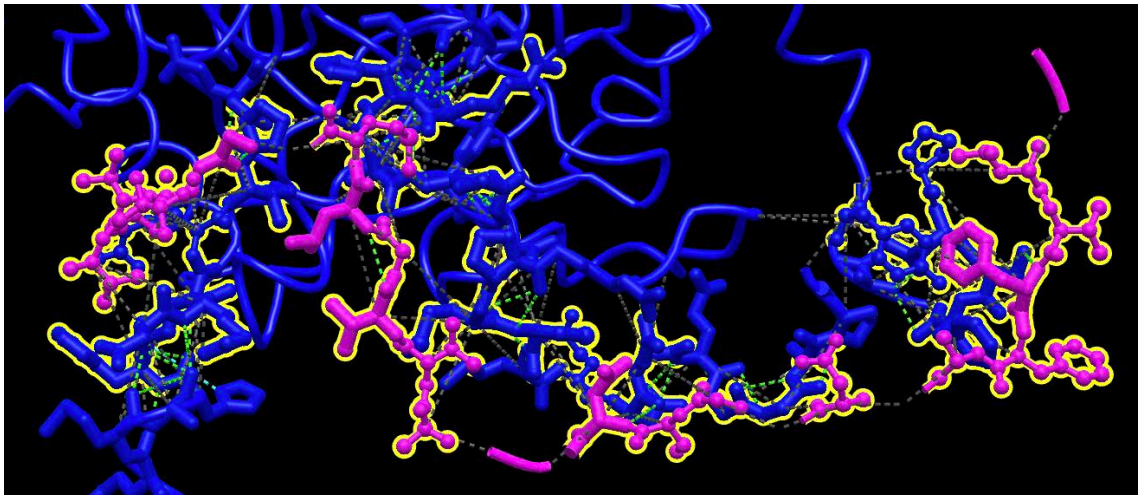

Within a given table the information given the entries detail which atom of a given amino acid in the chain for amyloid-ß interacts with which atom of a given amino acid in the chain for catalase. In this example the MET $stru.A:35@C and MET $stru.A:35@N are different atoms from the amyloid-ß methionine 35 residue. The ARG $stru.B:363@SD, HIS $stru.B:364@SD, GLY $stru.B:367@CE and PRO $stru.B:368@CE are different atoms from catalase residues arginine 363, histidine 364, glycine 367 and proline 368. The Min Distance (Å) represents the minimum distance between the atoms and the C-alpha Distance (Å) represents the distance between the C-alpha atoms of each amino acid. The identified residues can be compared with the BLAST search identified residues (see section 7e, page 39) and also the table of Molecular Recognition scoring table (see section 7b, page 31), in this example the amyloid-ß Met35 and catalase His364 would be predicted to interact based on the Molecular Recognition scoring but not the BLAST search (Milton *et al.,* 2001).



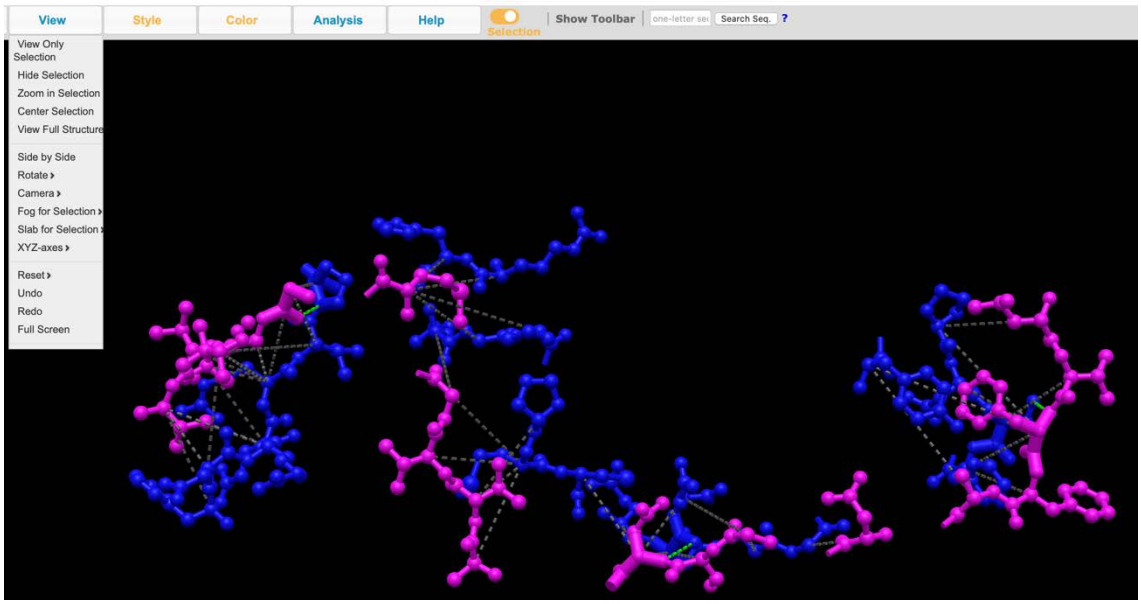

10(f) Diagrams of the interactions can also be manipulated and downloaded. By selecting the Style option for Proteins and or Side Chains the style of the selected components (in this case the interacting residues can be modified:



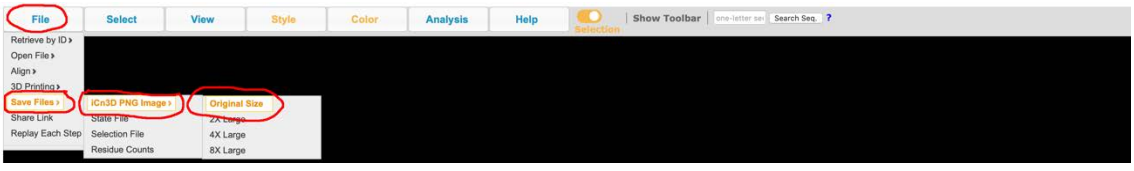
Resulting in this type of image:



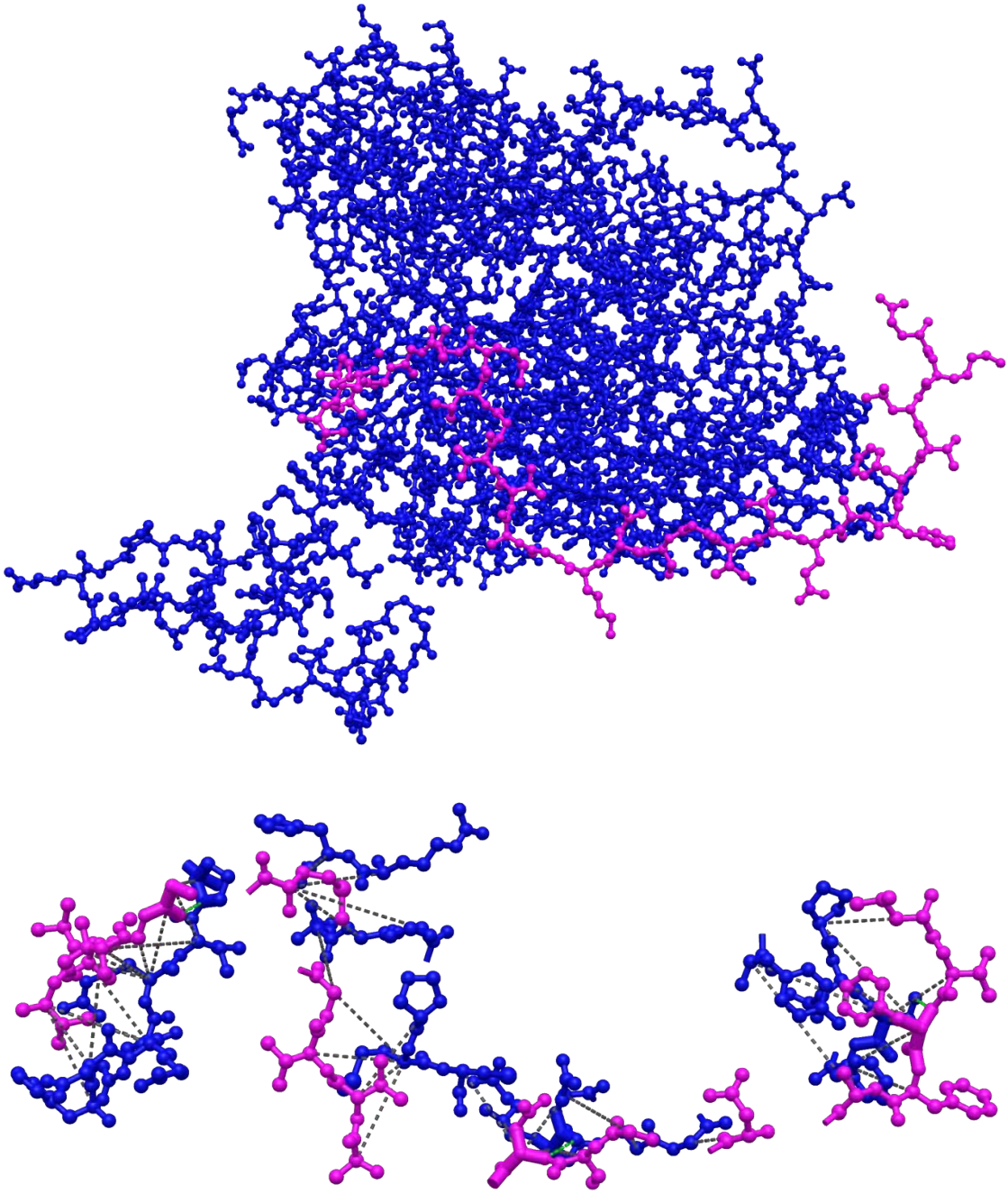
By selecting the View option and View Only Selection, the other parts of the structure are removed from the image.




Using the File, Save Files, iCn3D PNG Image options it is possible to download an image of the currently displayed screen.

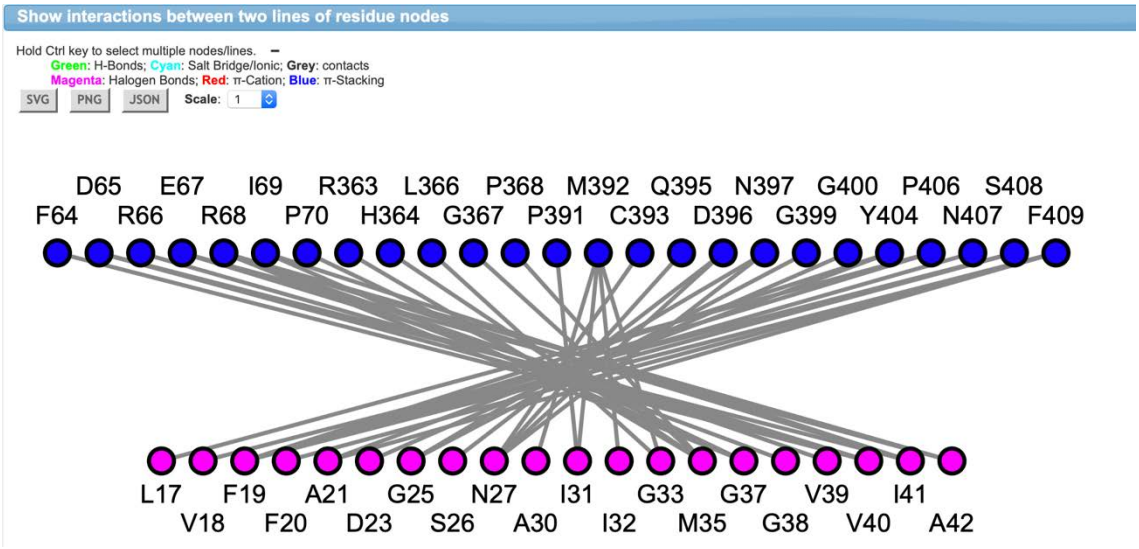



Examples of downloaded images from the Catalase (blue) Amyloid-ß (pink) interaction showing the full molecular structure plus interacting molecules of the structure:

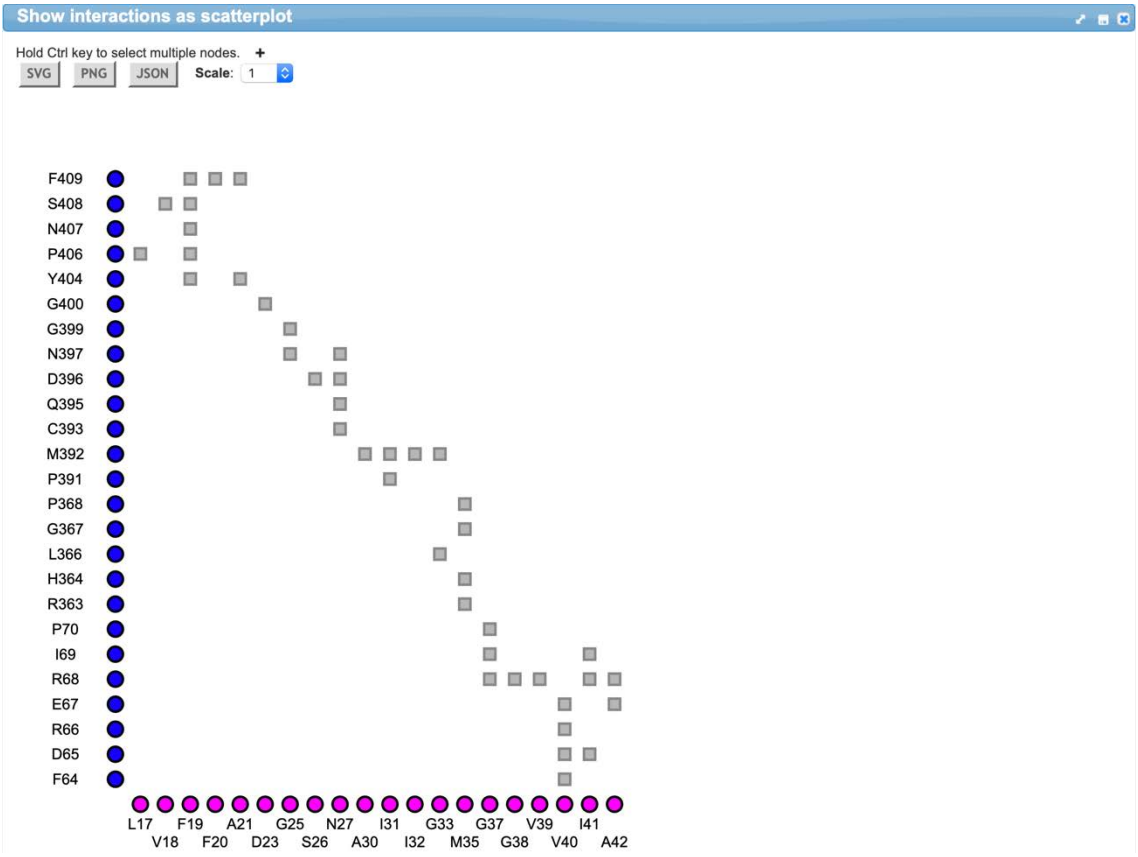


(10g) Other outputs included in the viewer are detailed in the Help Menu (see 10a, page 56 above). All of these features come with default settings, which can be modified as required. Links to each of these specific web pages can be downloaded as described in section 10c on page 58 above.

(i) 2D Interaction Network:

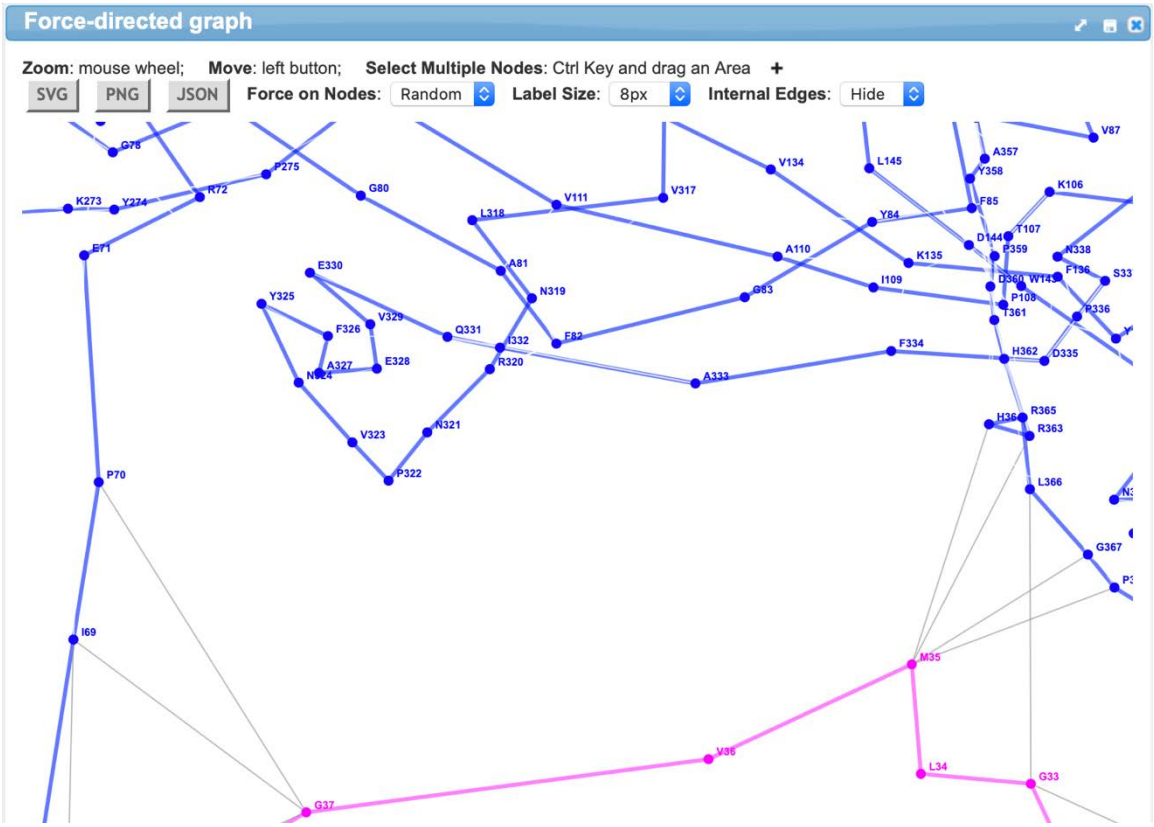



(ii) 2D Interaction Map:

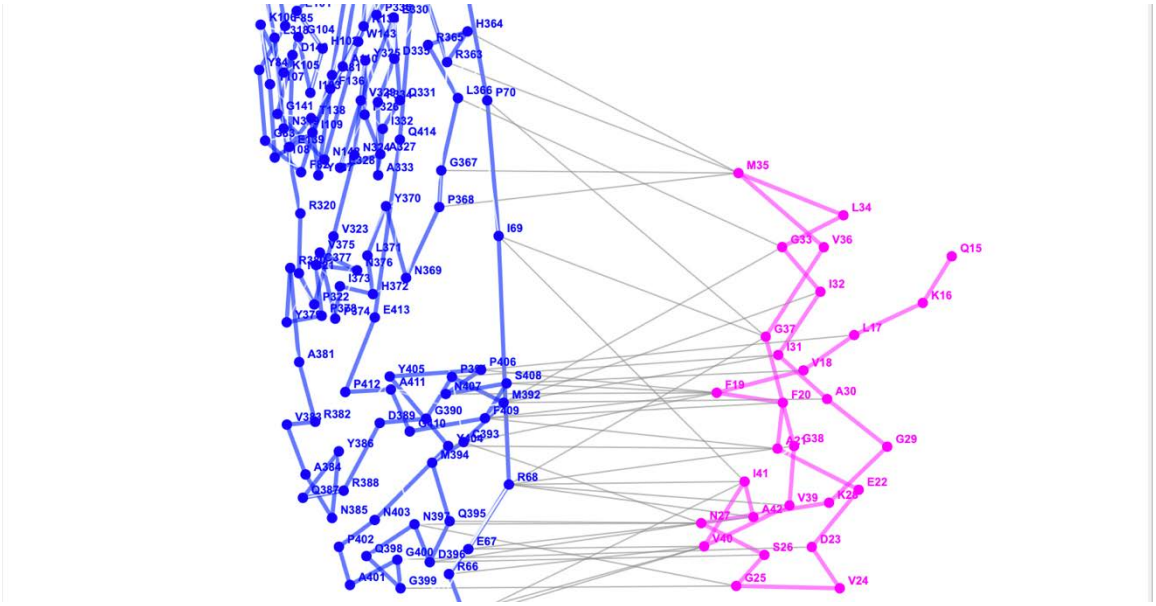


(iii) 2D Graph (Force-Directed), which by selecting the Force on Nodes options (circled in red on the Random version below) can be viewed in different formats:

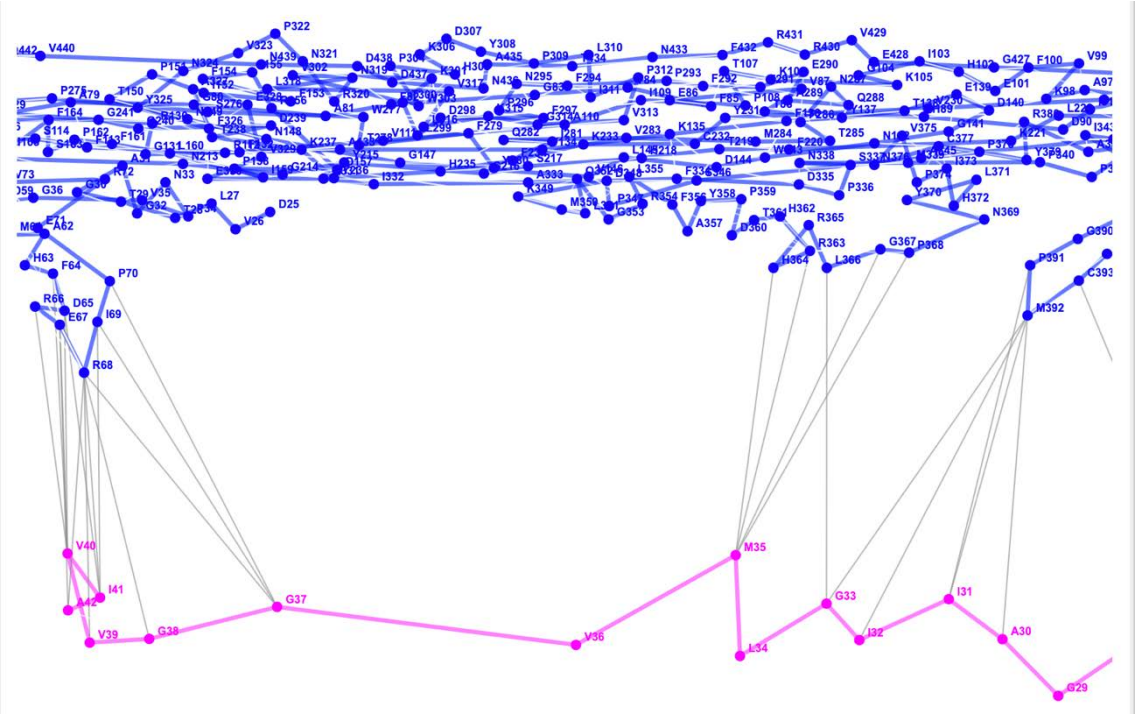
Random:



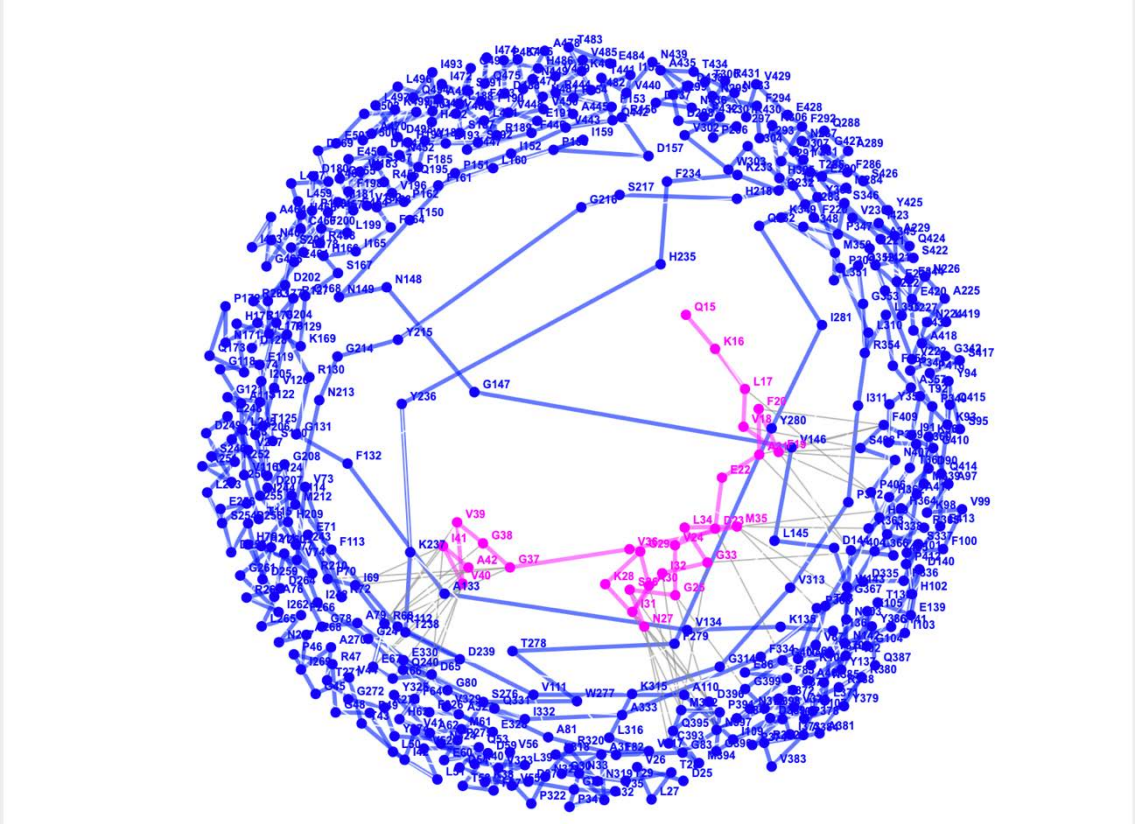

X-axis:



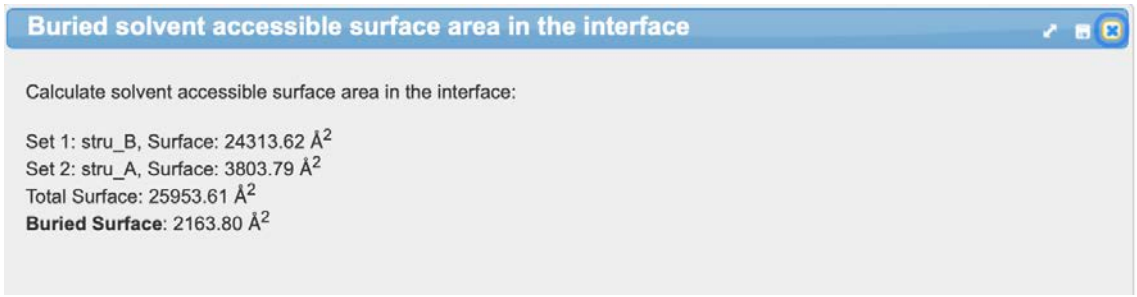
Y-axis:



Circle:

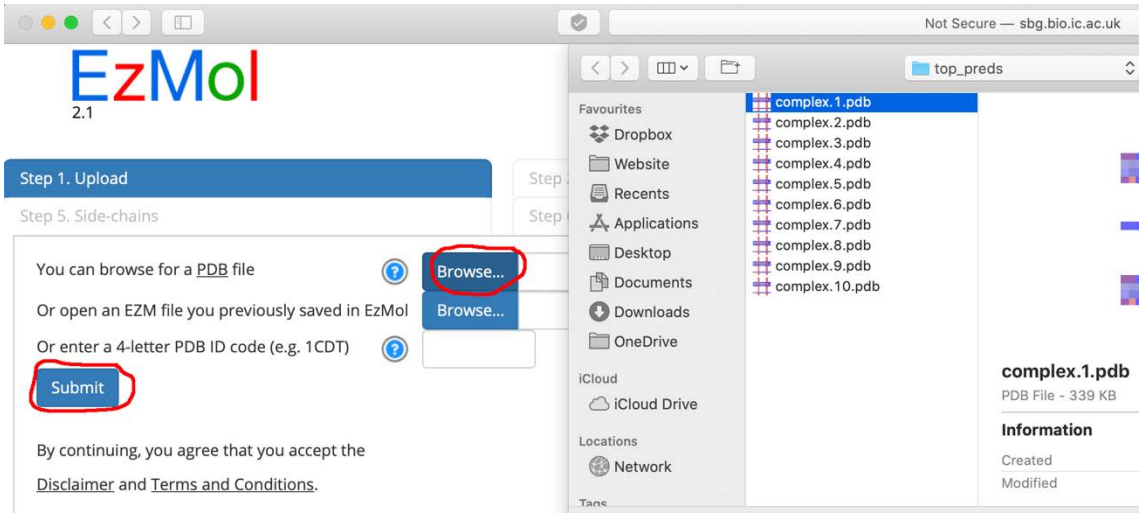


(iv) Buried Surface Area:

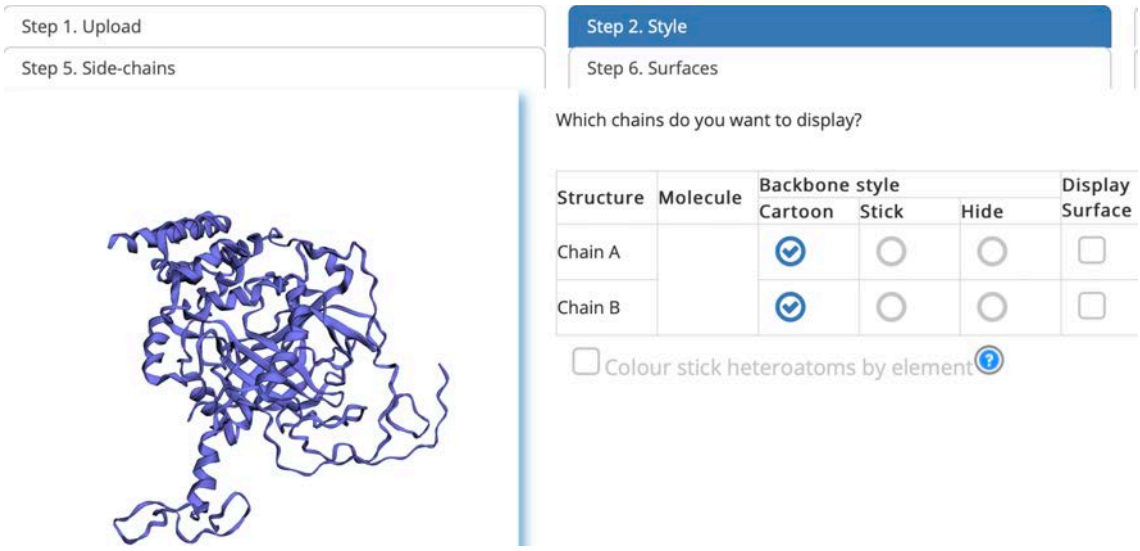


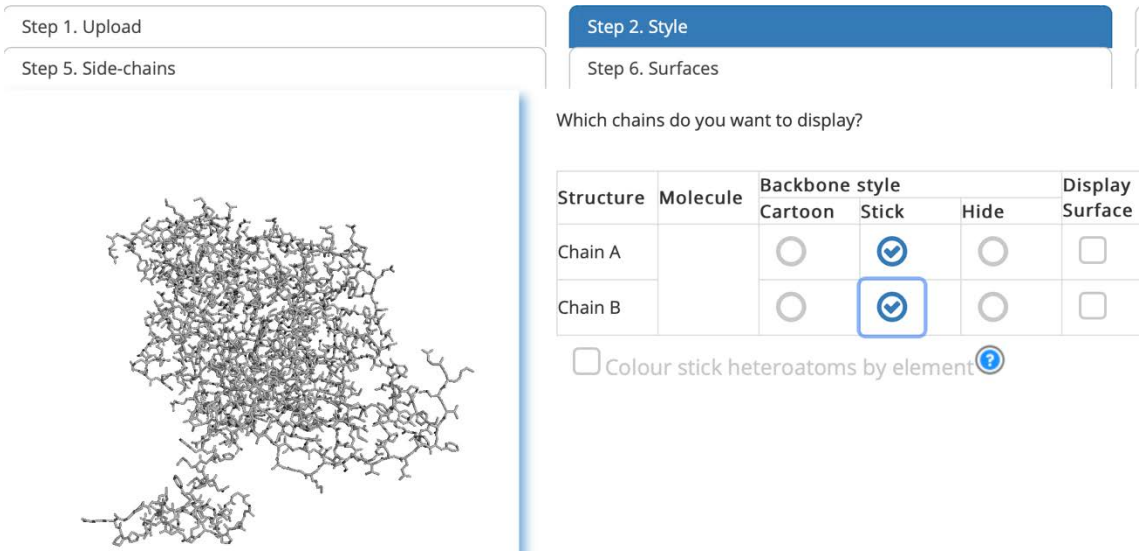
# 11: Protein-protein interaction images using EzMol

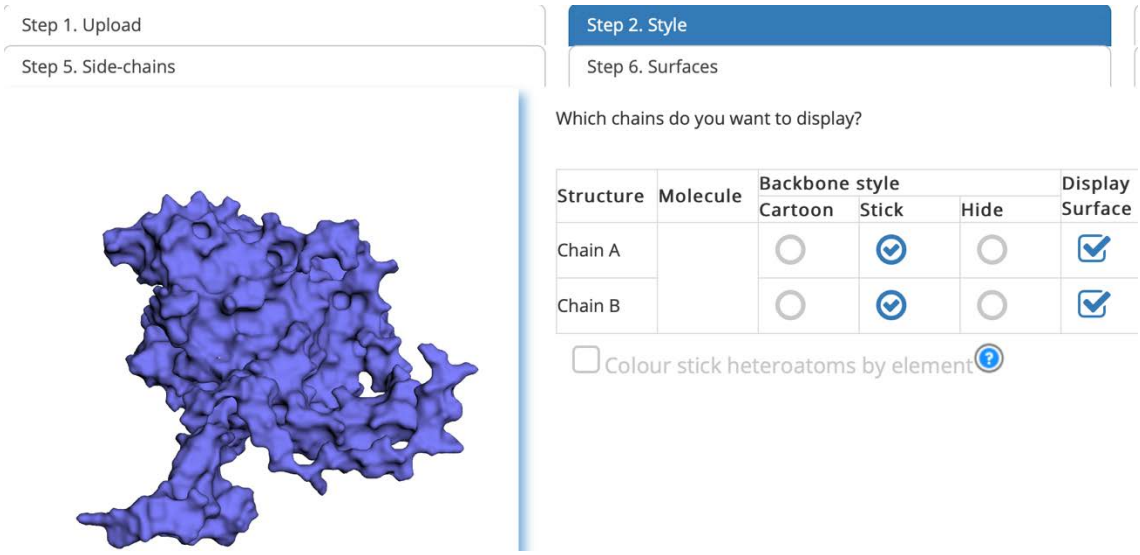
(11a) Image files can be created by uploading PDB files using the EzMol structure display (http://www.sbg.bio.ic.ac.uk/ezmol/) as described by Reynolds *et al.,* (2018). In Step 1 the PDB file is uploaded.



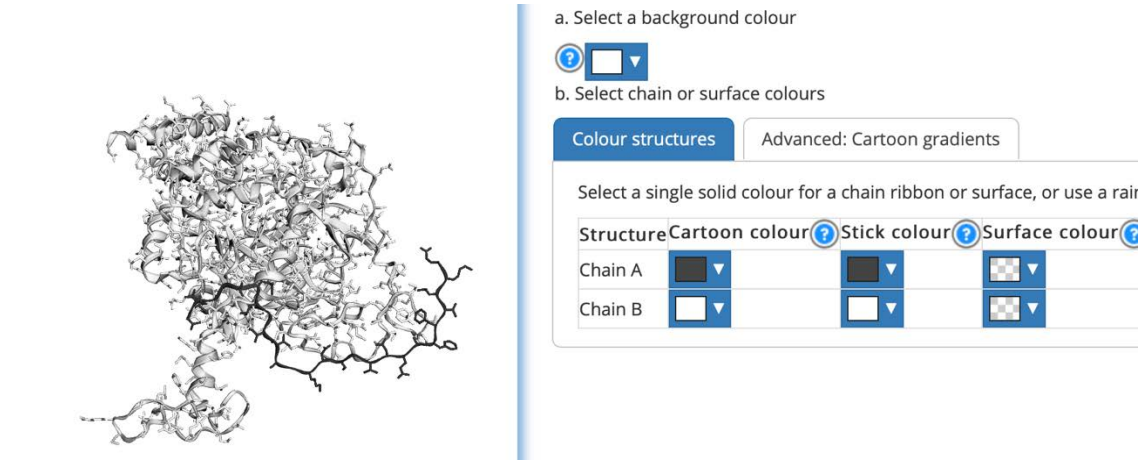

(11b) In Step 2 the type of display can be selected for each chain (in this example corresponding to amyloid-ß for chain A and catalase for chain B. The options are cartoon or stick with the ability to display the surface of the molecules.



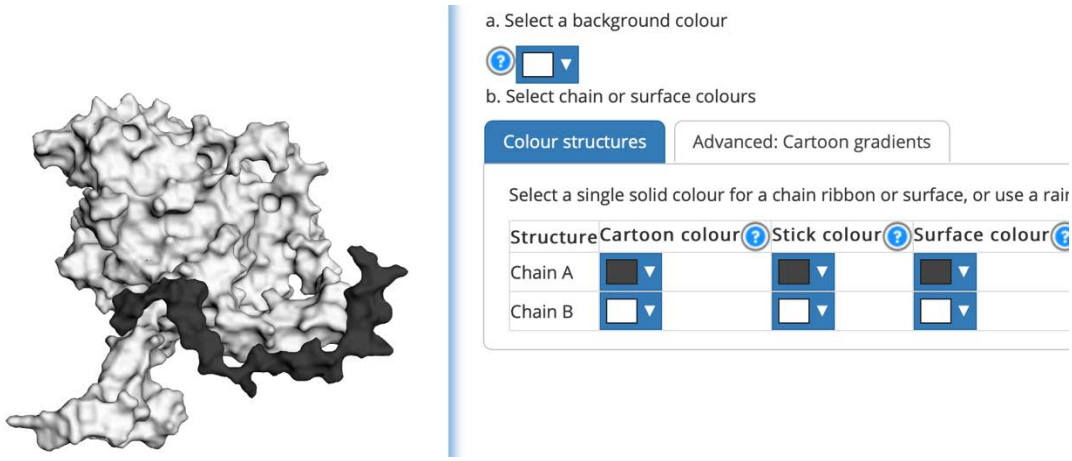


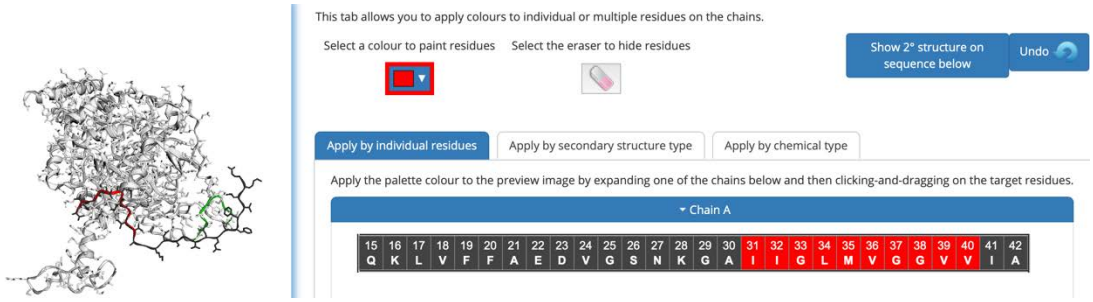

(11c) In Step 3 the colours of the displayed structures can be selected for each chain (in this example corresponding to amyloid-ß for chain A and catalase for chain B. The options allow colour selection for the whole of each chain. In the first example amyloid-ß is selected dark grey for both cartoon and stick colour and catalase is selected white.



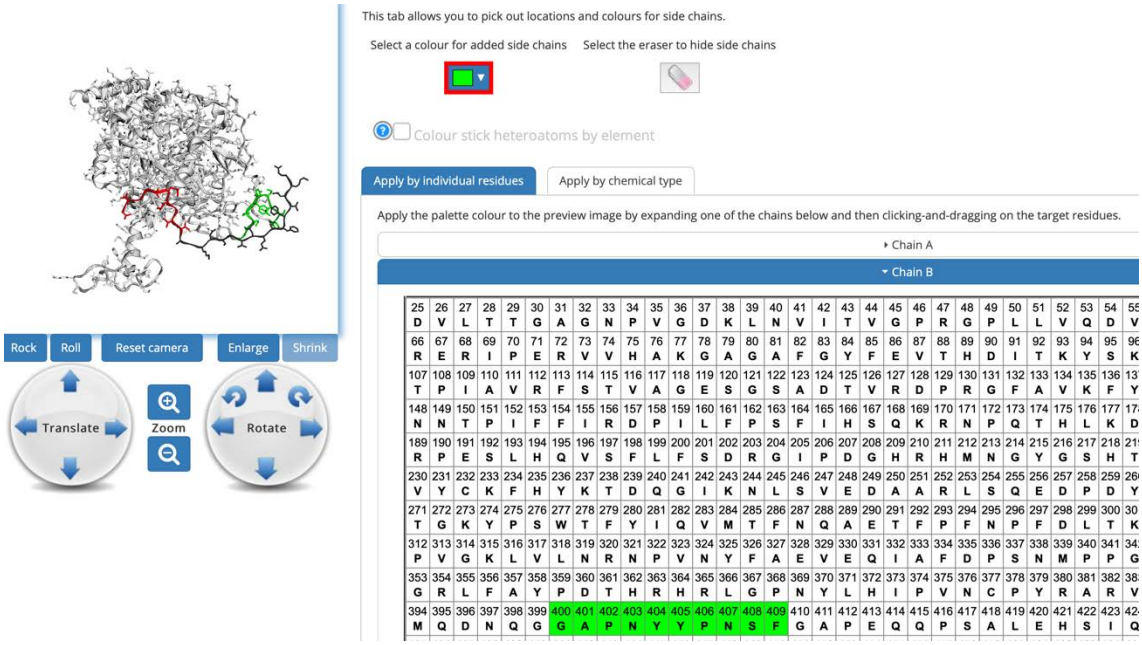
In the second example the surface structure is chosen with amyloid-ß selected as dark grey and catalase selected as white.



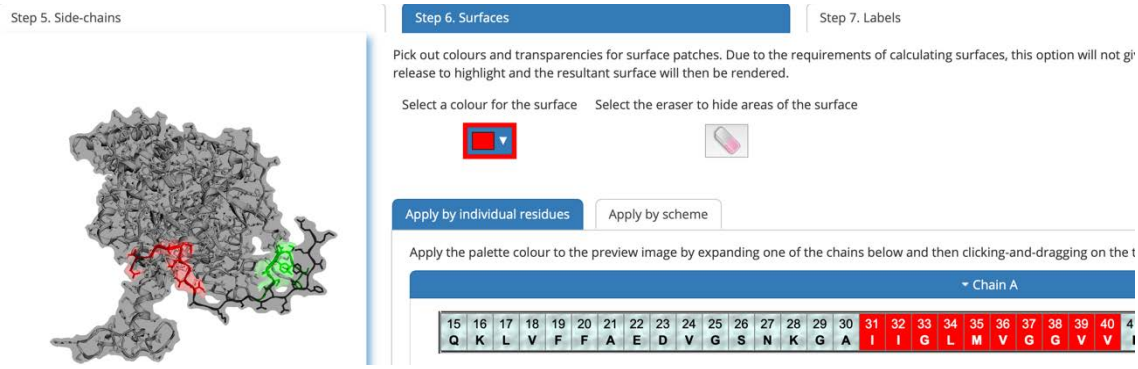

(11d) Step 4 allows selection of different colours for specific residues of the main chain and cartoon. In the example amyloid-ß 31-40 (IIGLMGGVV) have been selected as red and catalase 400-409 (GAPNYYPNSF) selected as green.



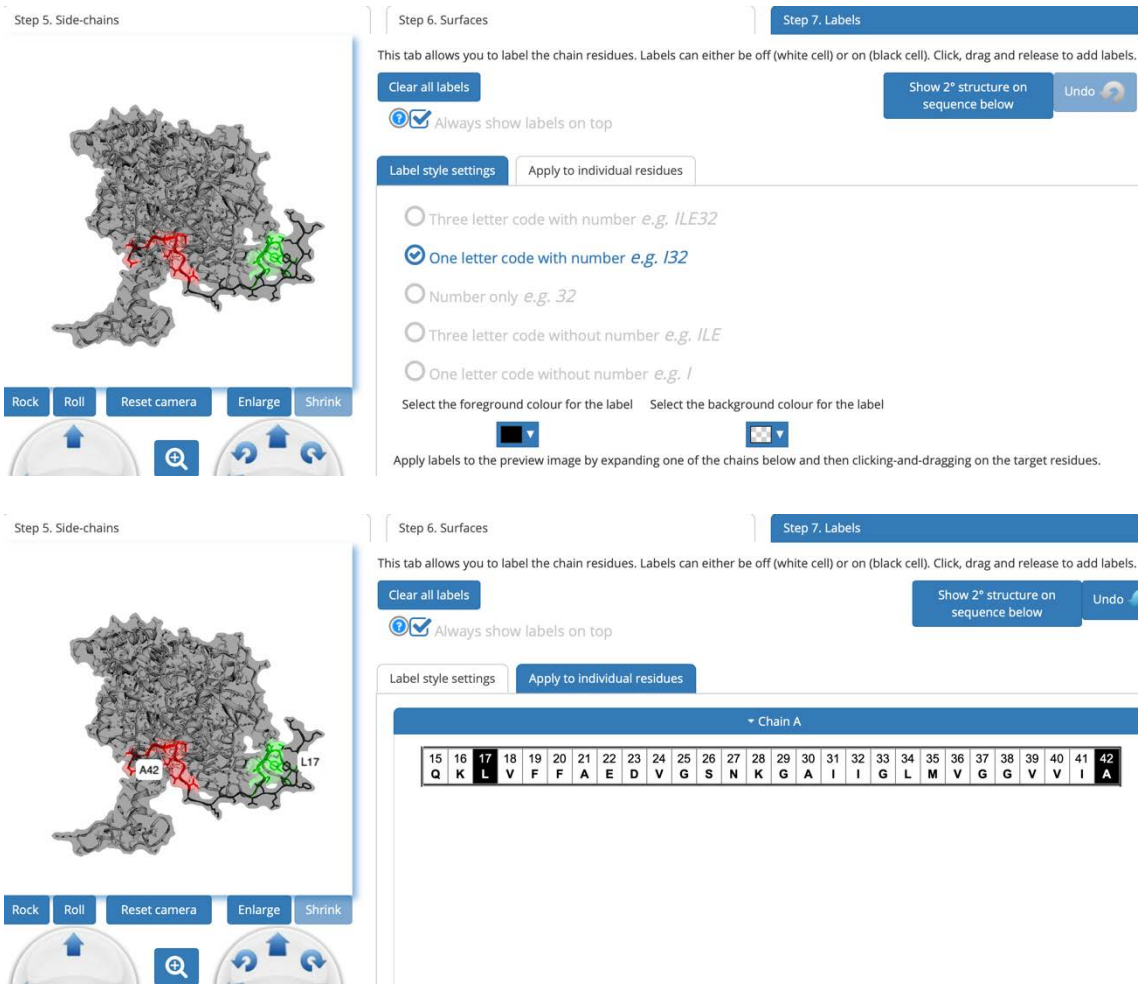

(11e) Step 5 allows the selection different colours for the specific side chains of residues.



(11f) Step 6 allows the selection different colours for the surface structure of residues; this is preferable to selecting surface colour in step 3 as it shows the chain detail within the model. In the example again the amyloid-ß 31-40 (IIGLMGGVV) have been selected as red and the catalase 400-409 (GAPNYYPNSF) selected as green.

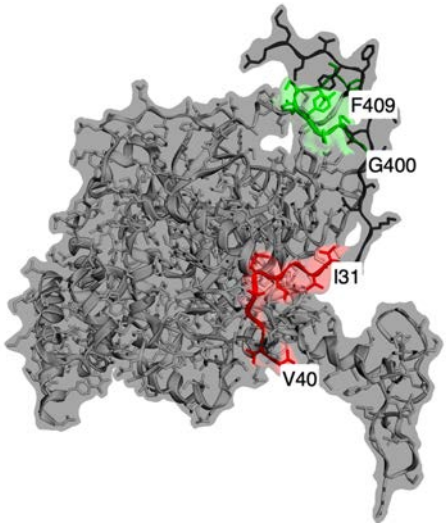



(11g) Step 7 allows the insertion of labels where the colour of the background and label together with the format applies to all the added labels.



(11h) The final Step 8 generates a rendered image that can then be downloaded as a png file. Using the controls below the image the model can be rotated to achieve the desired version for presentation.


Example of downloaded png file showing the interaction between amyloid-ß and catalase with specific residues from Milton *et al.,* 2001 labelled.

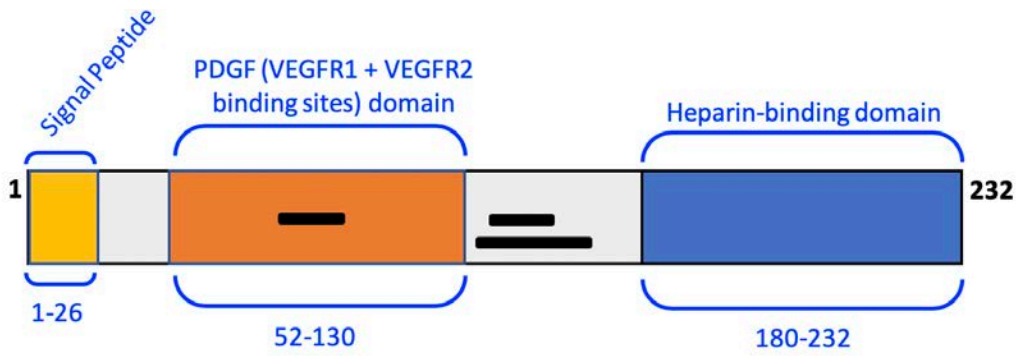


# 12: Interpretation of results

(12a) The target protein originally chosen (see Section 2 pages 5-10) will have a protein id (see Section 2(b) page 7) that can be searched to obtain further information from the NCBI Proteins (https://www.ncbi.nlm.nih.gov/protein/) or UniProt (https://www.uniprot.org) websites. The SNC sequence can also be BLAST searched (see Section 4(a) page 14) to obtain information about related proteins and isoforms of the target protein. Key features to identify for the target protein include modified residues (for example phosphorylated residues); residues that bind co-factors, ligands, allosteric ligands, substates or other proteins; regions linked to protein activity (for example the active sites of enzymes or ligand binding domains of receptors); regions that play a role in localisation (for example the extracellular, transmembrane and intracellular regions of receptors); regions where a protein undergoes post-translational cleavage (for example signal peptides or hormone pre- and pro- forms); regions with structural details (for example ß-turns).

(12b) These key features also need to be identified for each of potential binding proteins listed in the results tables (see Section 7(e) page 39 above) from the BLAST searches for alignments with AS35NC, AS35CN, AS53NC and AS53CN sequences.

(12c) A graphical map of potential interaction domains can be created from the data obtained and is particularly useful to look at the potential effects of interactions.




***A graphical view of the domains and interacting regions within VEGFA.*** *The black boxes within the chain represent the protein residues in which VEGFA binds to cubilin precursor, gametogenetin binding protein 2, and FGFR2 isoform 7 precursor. PDGF = platelet-derived growth factor.*

Similar graphical views can be prepared for each interacting protein, in this example the cubulin precursor, gametogenetin binding protein 2 and FGFR2 isoform 7 precursor.

(12d) Searches of publications using PubMed (https://pubmed.ncbi.nlm.nih.gov), Google Scholar (https://scholar.google.com) and Science Direct (https://www.sciencedirect.com) can be used to identify published links or processes that are linked to each protein.

(12e) Tissue localisation of the proteins is also very useful to help determine if an interaction is likely, this information may be provided in publications and can also be checked using the online Human Protein Atlas (https://www.proteinatlas.org).

(12f) From the interaction modelling results tables from section 10(e) (pages 60-61) it is possible to determine which amino acids are predicted to interact. This information can be combined with the information from the results tables (see Section 7(e) page 39 above) from the BLAST searches, for alignments with AS35NC, AS35CN, AS53NC and AS53CN sequences, to determine if the two techniques identify similar regions of proteins involved in interactions. Information about published interactions for each protein can be found at the IntAct Molecular Interaction Database (https://www.ebi.ac.uk/intact/) and structural information can also be found at the RSCB Protein Databank (https://www.rcsb.org). From these data sets it is possible to suggest which residues may interact based on the Bioinformatic computer predictions.

(12g) Images of the suggested interacting structures from section 10 (pages 56-67) and section 11 (pages 68-72) can be generated to illustrate the suggested interactions.

(12h) From these results practical experiments can be designed to prove or disprove the theoretical interactions identified using the Antisense Peptide Bioinformatics and Molecular Docking protocols.

(12i) The antisense peptide sequences can also be used to generate synthetic peptides for use in experimental settings.

# 13: References and websites


[1] Bioinformatics Protocols (https://www.neurodelta.uk/Protocols/)

[2] Biro, J.C. (2007) The Proteomic Code: a molecular recognition code for proteins. *Theor. Biol. Med. Model.*, **4**, 45. (https://doi.org/10.1186/1742-4682-4-45).

[3] Blalock, J.E. & Bost, K.L. (1986) Binding of peptides that are specified by complementary RNAs. *Biochem. J.,* **234(3)**, 679-683. (https://doi.org/10.1042/bj2340679).

[4] Blalock, J.E. & Bost, K.L. (1988) Ligand receptor characteristics of peptides encoded by complementary nucleic acids: implications for a molecular recognition code. *Recent Prog. Horm. Res.*, **44**, 199-222. (https://doi.org/10.1016/b978-0-12-571144-9.50011-9).

[5] Blalock, J.E. & Smith, E.M. (1984) Hydropathic anti-complementarity of amino acids based on the genetic code. *Biochem. Biophys. Res. Commun.*, **121(1)**, 203-207. (https://doi.org/10.1016/0006-291x(84)90707-1).

[6] Bost, K.L. & Blalock, J.E. (1989a) Complementary Peptides as Interactive Sites for Protein Binding. *Viral Immunol.,* **2(4)**, 229-238. (https://doi.org/10.1089/vim.1989.2.229).

[7] Bost, K.L. & Blalock, J.E. (1989b) Preparation and use of complementary peptides. *Methods Enzymol.*, **168**, 16-28. (https://doi.org/10.1016/0076-6879(89)68005-6).

[8] Bost, K.L., Smith, E.M. & Blalock, J.E. (1985) Similarity between the corticotropin (ACTH) receptor and a peptide encoded by an RNA that is complementary to ACTH mRNA. *Proc. Natl. Acad. Sci. U S A.*, **82(5)**, 1372-1375. (https://doi.org/10.1073/pnas.82.5.1372).

[9] Capone, G., De Marinis, A., Simone, S., Kusalik, A. & Kanduc, D. (2008) Mapping the human proteome for non-redundant peptide islands. *Amino Acids.*, **35(1)**, 209-216. (https://doi.org/10.1007/s00726-007-0563-7).

[9a] Chen, R., & Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, **47(3)**, 281–294. (https://doi.org/10.1002/prot.10092).

[10] Chilumuri, A., Odell, M. & Milton, N.G. (2013a) Benzothiazole aniline tetra(ethylene glycol) and 3-amino-1,2,4-triazole inhibit neuroprotection against amyloid peptides by catalase overexpression in vitro. *ACS Chem. Neurosci.*, **4(11)**, 1501-1512. (https://doi.org/10.1021/cn400146a).

[11] Chilumuri, A., Ashioti, M., Nercessian, A. N., & Milton, N. G. (2013b). Immunolocalization of Kisspeptin Associated with Amyloid-β Deposits in the Pons of an Alzheimer's Disease Patient. *Journal of neurodegenerative diseases*, *2013*, 879710. (https://doi.org/10.1155/2013/879710).

[12] Clarke, B.L. & Blalock, J.E. (1990) Steroidogenic activity of a peptide specified by the reversed sequence of corticotropin mRNA. *Proc. Natl. Acad. Sci. U S A.*, **87(24)**, 9708-9711. (https://doi.org/10.1073/pnas.87.24.9708).

[13] EzMol structure display (http://www.sbg.bio.ic.ac.uk/ezmol/)

[14] Fassina, G., Roller, P.P., Olson, A.D., Thorgeirsson, S.S. & Omichinski, J.G. (1989) Recognition properties of peptides hydropathically complementary to residues 356-375 of the c-raf protein. *J. Biol. Chem.*, **264(19)**, 11252-11257. (https://www.jbc.org/content/264/19/11252.long)

[15] Google Scholar (https://scholar.google.com)

[16] Hardison, M.T. & Blalock, J.E. (2012) Molecular recognition theory and sense-antisense interaction: therapeutic applications in autoimmunity. *Front. Biosci. (Elite Ed).*, **4**, 1864-1870. (https://doi.org/10.2741/508).

[17] Heal, J.R., Roberts, G.W., Raynes, J.G., Bhakoo, A. & Miller A.D. (2002) Specific interactions between sense and complementary peptides: the basis for the proteomic code. *Chembiochem.*, **3(2-3)**, 136-151. (https://doi.org/10.1002/1439-7633(20020301)3:2/3%3c136::aid-cbic136%3e3.0.co;2-7).

[18] iCn3D struct viewer (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html)

[19] Illingworth, C.J., Chintipalli, S.V., Serapian, S.A., Miller, A.D., Veverka, V., Carr, M.D. & Reynolds, C.A. (2012) The statistical significance of selected sense-antisense peptide interactions. *J. Comput. Chem.*, **33(16)**, 1440-1447. (https://doi.org/10.1002/jcc.22977).

[20] IntAct Molecular Interaction Database (https://www.ebi.ac.uk/intact/)

[21] I-Tasser PDB Prediction (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)

[22] Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157(1)**, 105-132. (https://doi.org/10.1016/0022-2836(82)90515-0).

[23] McGuire, K.L. & Holmes, D.S. (2005) Role of complementary proteins in autoimmunity: an old idea re-emerges with new twists. *Trends Immunol.*, **26(7)**, 367-372. (https://doi.org/10.1016/j.it.2005.05.001).

[24] Miller, A.D. (2015) Sense-antisense (complementary) peptide interactions and the proteomic code; potential opportunities in biology and pharmaceutical science. *Expert Opin. Biol. Ther.*, **15(2)**, 245-267. (https://doi.org/10.1517/14712598.2015.983069).

[25] Milton, N.G.N., Mayor, N.P. & Rawlinson, J. (2001) Identification of amyloid-beta binding sites using an antisense peptide approach. *Neuroreport*, **12(11)**, 2561-2566. (https://doi.org/10.1097/00001756-200108080-00054).

[26] Milton, N.G.N. (2006) Anti-sense Peptides. In *Cell Biology Protocols*, Eds D. Rickwood, J. Graham & J.R. Harris, Wiley, London, pp 353-358. (https://doi.org/10.1002/0470033487.ch6).

[27] Model Archive (https://www.modelarchive.org)

[28] Mulchahey, J.J., Neill, J.D., Dion, L.D., Bost, K.L. & Blalock, J.E. (1986) Antibodies to the binding site of the receptor for luteinizing hormone-releasing hormone (LHRH): generation with a synthetic decapeptide encoded by an RNA complementary to LHRH mRNA. *Proc. Natl. Acad. Sci. U S A.*, **83(24)**, 9714-9718. (https://doi.org/10.1073/pnas.83.24.9714).

[29] NCBI Nucleotide Database (https://www.ncbi.nlm.nih.gov/nucleotide/)

[30] NCBI Protein Database (https://www.ncbi.nlm.nih.gov/protein/)

[31] NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

[32] PDBx/mmCIF conversion (https://mmcif.pdbj.org/converter/index.php?l=en)

[33] PDB chain renaming (http://www.canoz.com/sdh/renamepdbchain.pl)

[34] PDB information (https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure)

[35] Pearson W. R. (2013). Selecting the Right Similarity-Scoring Matrix. *Curr. Prot. Bioinf.*, **43**, 3.5.1–3.5.9. (https://doi.org/10.1002/0471250953.bi0305s43).

[36] Pierce, B. G., Hourai, Y., & Weng, Z. (2011). Accelerating protein docking in ZDock using an advanced 3D convolution library. *PloS One*, **6(9)**, e24657. (http://doi:10.1371/journal.pone.0024657).

[37] Pierce, B.G., Wiehe, K., Hwang, H., Kim, B.H., Vreven, T. & Weng, Z. (2014) ZDock server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics.*, **30(12)**, 1771-1773. (https://doi.org/10.1093/bioinformatics/btu097).

[38] PubMed (https://pubmed.ncbi.nlm.nih.gov)

[39] Pullen, J.R., Dalmaris, J., Serapian, S.A. & Miller, A.D. (2013) Assessing the preferred solution conformation of an interacting sense-antisense (complementary) peptide pair. *Bioorg. Med. Chem. Lett.*, **23(2)**, 496-502. (https://doi.org/10.1016/j.bmcl.2012.11.038).

[40] Putnam, C.D., Arvai, A.S., Bourne, Y., & Tainer, J. A. (2000). Active and inhibited human catalase structures: ligand and NADPH binding and catalytic mechanism. *Journal of molecular biology*, **296(1)**, 295–309. (https://doi.org/10.1006/jmbi.1999.3458)

[41] Python (https://www.python.org)

[42] Python complier (https://trinket.io/python3)

[43] Reynolds, C. R., Islam, S. A., & Sternberg, M. (2018). EzMol: A Web Server Wizard for the Rapid Visualization and Image Production of Protein and Nucleic Acid Structures. *Journal of molecular biology*, **430(15)**, 2244–2248. (https://doi.org/10.1016/j.jmb.2018.01.013).

[44] Root-Bernstein R.S. & Holsworth, D.D. (1998) Antisense peptides: a critical mini-review. *J. Theor. Biol.*, **190(2)**, 107-119.

(https://doi.org/10.1006/jtbi.1997.0544).

[45] RSCB Protein Databank (https://www.rcsb.org)

[46] Science Direct (https://www.sciencedirect.com)

[47] Siemion, I.Z., Cebrat, M. & Kluczyk, A. (2004) The problem of amino acid complementarity and antisense peptides. *Curr. Protein Pept. Sci.*, **5(6)**, 507-527. (https://doi.org/10.2174/1389203043379413).

[48] Štambuk, N., Konjevoda, P., Boban-Blagaić, A. & Pokrić, B. (2005) Molecular Recognition Theory of the complementary (antisense) peptide interactions. *Theory Biosci.*, **123(4)**, 265-275. (https://doi.org/10.1016/j.thbio.2005.02.001).

[49] Štambuk, N., Konjevoda, P., Turčić, P., Kövér, K., Kujundžić, R.N., Manojlović, Z. & Gabričević, M. (2018). Genetic coding algorithm for sense and antisense peptide interactions. *Biosystems.*, **164**, 199-216. (https://doi.org/10.1016/j.biosystems.2017.10.009).

[50] Štambuk, N., Konjevoda, P., Turčić, P., Šošić, H., Aralica, G., Babić, D., Seiwerth, S., Kaštelan, Ž., Kujundžić, R.N., Wardega, P., Žutelija, J.B., Gračanin, A.G. & Gabričević, M. (2019). Targeting Tumor Markers with Antisense Peptides: An Example of Human Prostate Specific Antigen. *Int. J. Mol. Sci.*, **20(9)**, 2090. (https://doi.org/10.3390/ijms20092090).

[50a] Štambuk, N., Konjevoda, P., & Pavan, J. (2021). Antisense Peptide Technology for Diagnostic Tests and Bioengineering Research. *International journal of molecular sciences*, **22(17)**, 9106. (https://doi.org/10.3390/ijms22179106).

[51] The Human Protein Atlas (https://www.proteinatlas.org)

[52] UniProt (https://www.uniprot.org)

[53] Upsidedown text (http://www.upsidedowntext.com/)

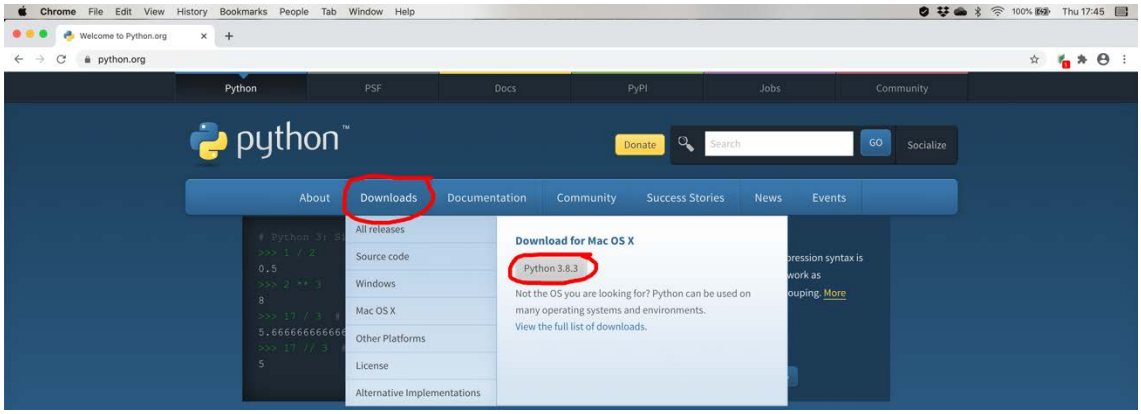
[54] Wang, J., Youkharibache, P., Zhang, D., Lanczycki, C. J., Geer, R. C., Madej, T., Phan, L., Ward, M., Lu, S., Marchler, G. H., Wang, Y., Bryant, S. H., Geer, L. Y., & Marchler-Bauer, A. (2020). iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics (Oxford, England)*, **36(1)**, 131–135. (https://doi.org/10.1093/bioinformatics/btz502).

[55] ZDock server (http://zdock.umassmed.edu)

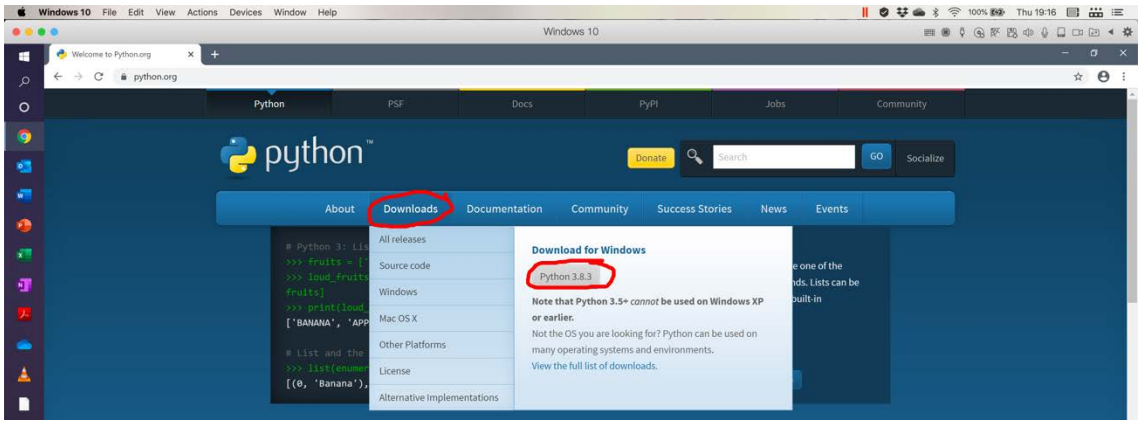
# 14: Appendix 1 - Installing Python

(14a) This protocol uses Python 3 software to run a program. Recommended to download the latest stable version, the program used has been tested on versions 3.6.0 and above (current stable version is 3.9.4). If Python cannot be installed either use an online version of Python (Section 3 above, pages 11-13) or an alternative method to generate antisense peptides manually is detailed in Section 15 (pages 83-86 below).

(14b) On a Mac to download Python go to https://www.python.org and follow instructions:

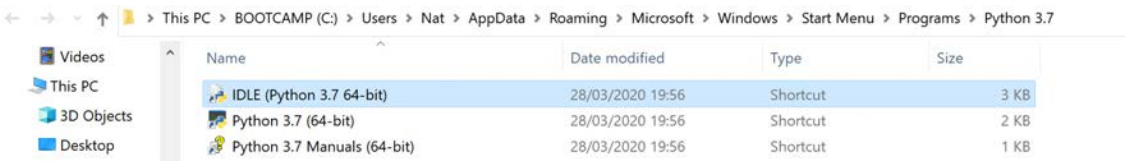



On a PC to download Python go to https://www.python.org and follow instructions:




(14c) Once installed on a Mac there will be a Python folder in the Applications (Mac), to open double click on the IDLE icon (highlighted in blue):

On a PC the Python folder should be on the C drive (PC) in the Programs folder and to open double click on the IDLE icon (highlighted in blue):

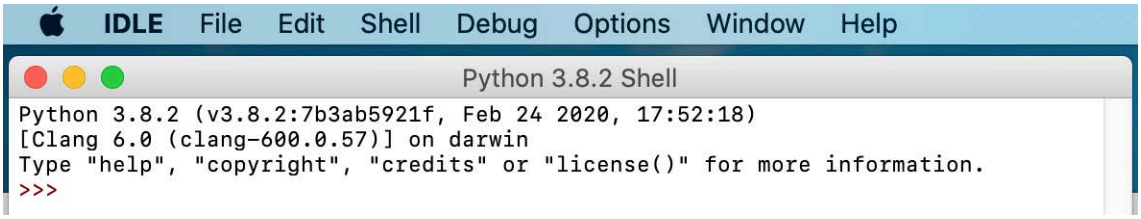


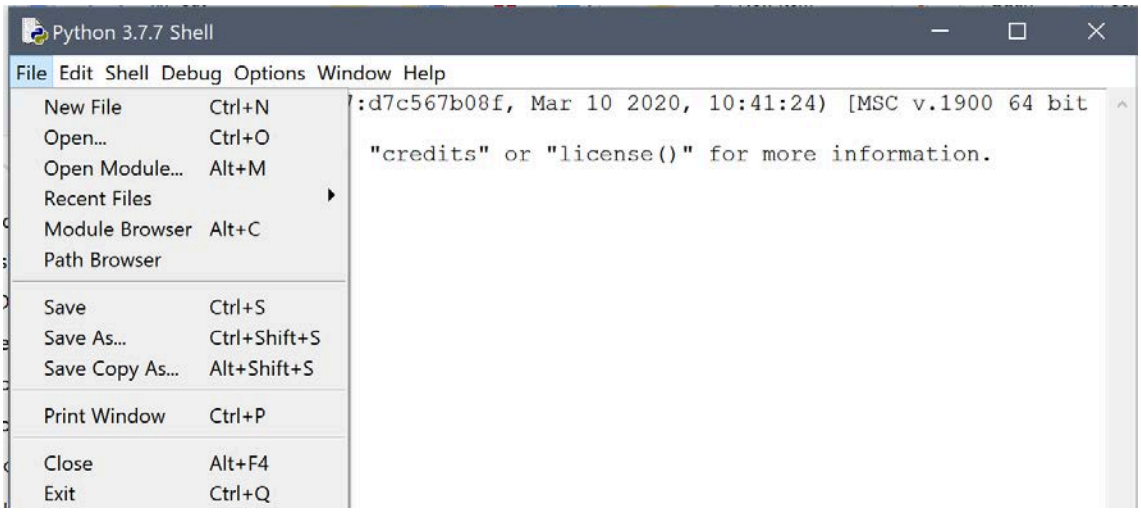

(14d) Download the Python script (AntisensePeptide.py) file from Antisense-Peptide.py (available from as a either Python script https://www.neurodelta.uk/resources/BioinformaticsProtocolScript.py save to a suitable folder on the hard drive. A copy the complete text is shown below:

```
##Original code by Jonathan C Goulding, Adapted to Py3 and Modified by
Harrison R S Milton, based on Milton, N.G.N. (2006) Anti-sense Peptides.
Protocols 6.39 In Cell Biology Protocols, Eds D. Rickwood, J. Graham &
J.R. Harris, Wiley, London, pp 353-358. ? J.C. Goulding, H.R.S. Milton &
N.G.N. Milton; School of Clinical & Applied Sciences, Leeds Beckett
University; Neurodelta Ltd.
g=input("Input name: ")
s=input("Input coding mRNA: ")
i=s.replace(' ', '').replace('0', '').replace('1', '').replace('2',
'').replace('3', '').replace('4', '').replace('5', '').replace('6',
'').replace('7', '').replace('8', '').replace('9', '').replace('A',
'a').replace('C', 'c').replace('G', 'g').replace('T', 't').replace('U',
't').replace('u', 't').replace('\n', '')

def breakdown(data):
    array=[]
    for i in range(0,len(data),3):
        if (i+3>len(data)):
            upper = len(data)
        else:
            upper =i+3
        seq=data[i:upper]
        try:
            array.append(amino_acids[seq])
        except KeyError :
            array.append('unknown')
    return array

def flip(x):
  return x[::-1]

amino_acids =
{'aaa':'K','aac':'N','aag':'K','aat':'N','aca':'T','acc':'T','acg':'T','a
ct':'T','aga':'R','agc':'S','agg':'R','agt':'S','ata':'I','atc':'I','atg'
:'M','att':'I','caa':'Q','cac':'H','cag':'Q','cat':'H','cca':'P','ccc':'P
','ccg':'P','cct':'P','cga':'R','cgc':'R','cgg':'R','cgt':'R','cta':'L','
ctc':'L','ctg':'L','ctt':'L','gaa':'E','gac':'D','gag':'E','gat':'D','gca
':'A','gcc':'A','gcg':'A','gct':'A','gga':'G','ggc':'G','ggg':'G','ggt':'
G','gta':'V','gtc':'V','gtg':'V','gtt':'V','tac':'Y','tat':'Y','tca':'S',
'tcc':'S','tcg':'S','tct':'S','tgc':'C','tgg':'W','tgt':'C','tta':'L','tt
c':'F','ttg':'L','ttt':'F','taa':'*','tga':'*','tag':'*'}
output=breakdown(i)
combined=''
for acid in output:
    combined =combined+acid
```

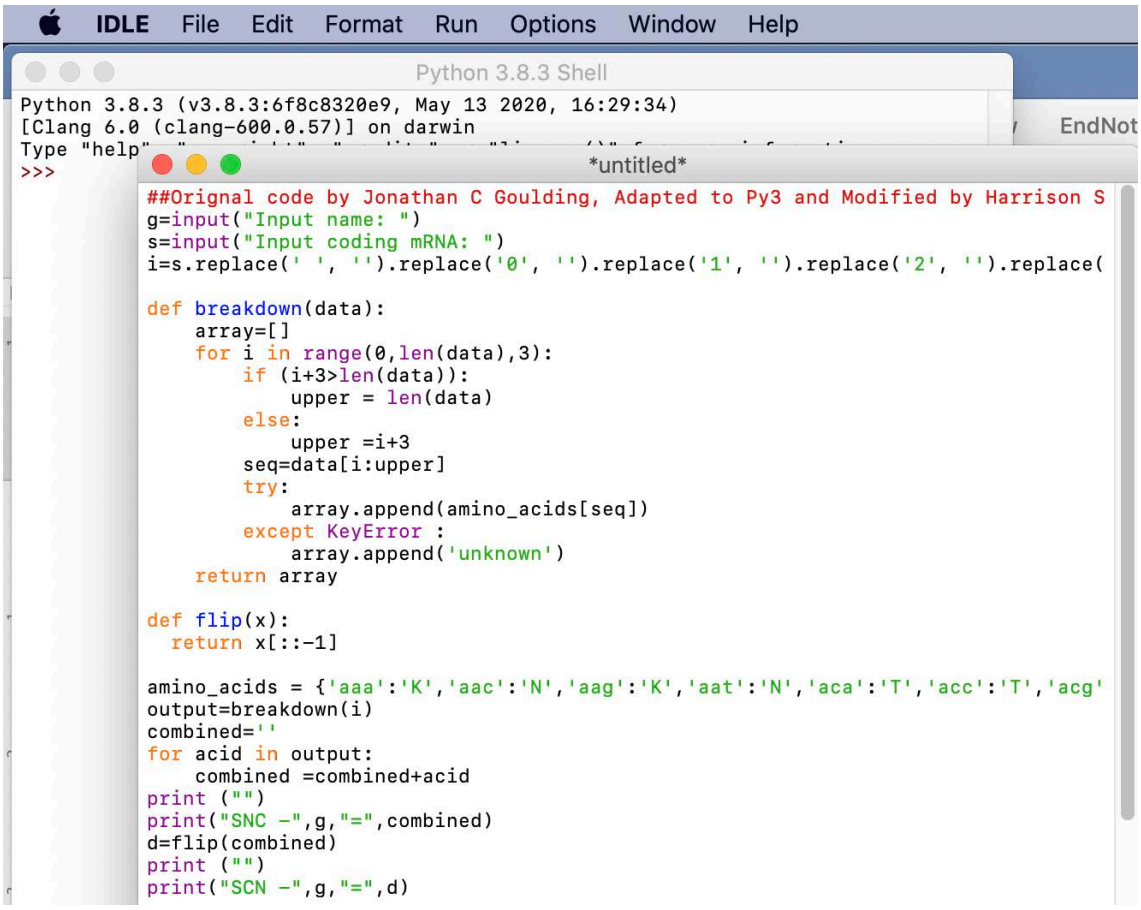
```
print ("")
print("SNC -",g,"=",combined)
d=flip(combined)
print ("")
print("SCN -",g,"=",d)

amino_acids =
{'aaa':'F','aac':'L','aag':'F','aat':'L','aca':'C','acc':'W','acg':'C','a
ct':'X','aga':'S','agc':'S','agg':'S','agt':'S','ata':'Y','atc':'X','atg'
:'Y','att':'X','caa':'V','cac':'V','cag':'V','cat':'V','cca':'G','ccc':'G
','ccg':'G','cct':'G','cga':'A','cgc':'A','cgg':'A','cgt':'A','cta':'D','
ctc':'E','ctg':'D','ctt':'E','gaa':'L','gac':'L','gag':'L','gat':'L','gca
':'R','gcc':'R','gcg':'R','gct':'R','gga':'P','ggc':'P','ggg':'P','ggt':'
P','gta':'H','gtc':'Q','gtg':'H','gtt':'Q','tac':'M','tat':'I','tca':'S',
'tcc':'R','tcg':'S','tct':'R','tgc':'T','tgg':'T','tgt':'T','tta':'N','tt
c':'K','ttg':'N','ttt':'K','taa':'*','tga':'*','tag':'*'}
output=breakdown(i)
combined=''
for acid in output:
    combined =combined+acid
print ("")
print("AS35NC -",g,"=",combined)
e=flip(combined)
print ("")
print("AS35CN -",g,"=",e)

amino_acids =
{'aaa':'F','aac':'V','aag':'L','aat':'I','aca':'C','acc':'G','acg':'R','a
ct':'S','aga':'S','agc':'A','agg':'P','agt':'T','ata':'Y','atc':'D','atg'
:'H','att':'N','caa':'L','cac':'V','cag':'L','cat':'M','cca':'W','ccc':'G
','ccg':'R','cct':'R','cga':'S','cgc':'A','cgg':'P','cgt':'T','cta':'X','
ctc':'E','ctg':'Q','ctt':'K','gaa':'F','gac':'V','gag':'L','gat':'I','gca
':'C','gcc':'G','gcg':'R','gct':'S','gga':'S','ggc':'A','ggg':'P','ggt':'
T','gta':'Y','gtc':'D','gtg':'H','gtt':'N','tac':'V','tat':'I','tca':'X',
'tcc':'G','tcg':'R','tct':'R','tgc':'A','tgg':'P','tgt':'T','tta':'X','tt
c':'E','ttg':'Q','ttt':'K','taa':'*','tga':'*','tag':'*'}
output=breakdown(i)
combined=''
for acid in output:
    combined =combined+acid
print ("")
print("AS53NC -",g,"=",combined)
d=flip(combined)
print ("")
print("AS53CN -",g,"=",d)
```

(14e) Alternatively in the Python Shell after clicking in File will give a pull down on which the New File (Ctrl+N) can be selected:






(14f) Download the Python script (AntisensePeptide.docx) file from https://www.neurodelta.uk/Protocols/. Open the document in Microsoft Word and Copy the complete text. Paste into the blank untitled window on Python:

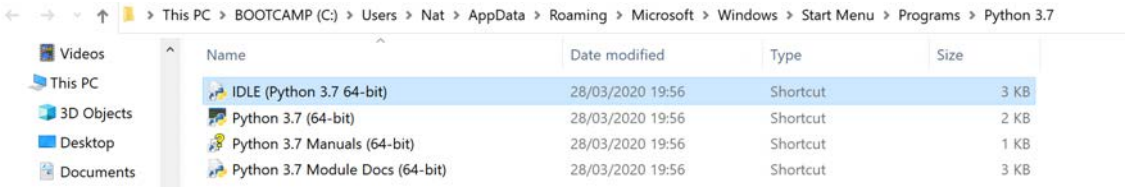



Then save the file on the Hard drive in a suitable folder as AntisensePeptide.py using the file Save (Ctrl+S) command.

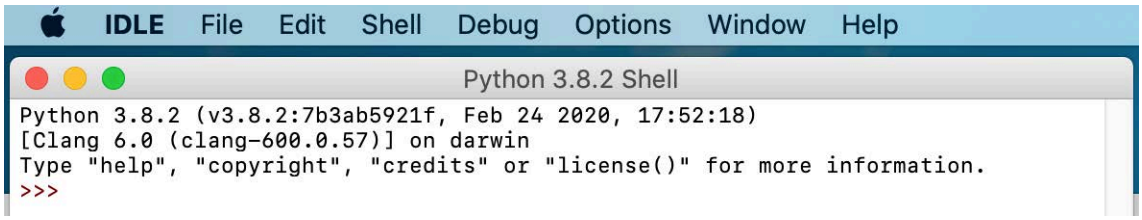
# 15: Appendix 2 - Python antisense peptide generation

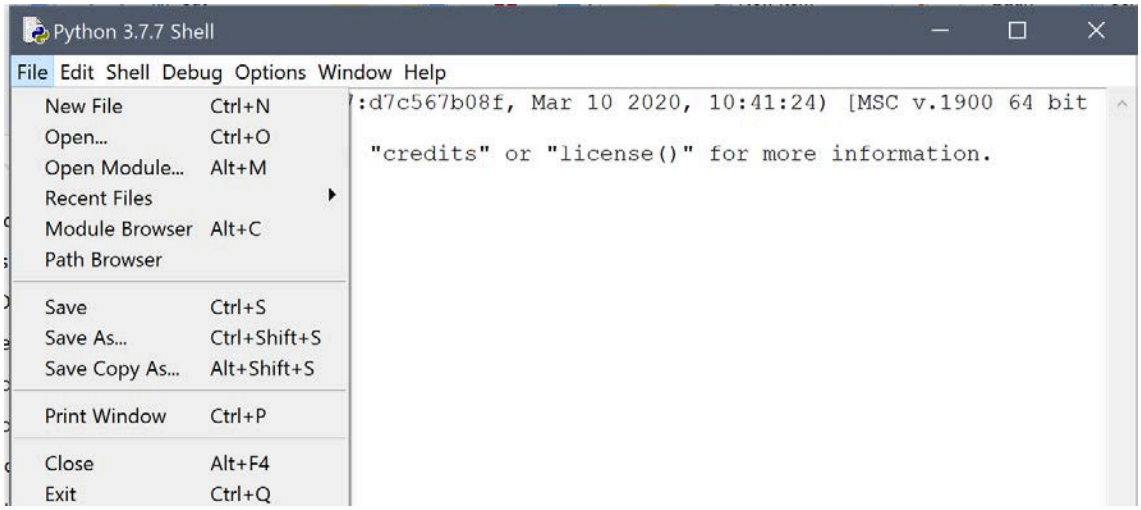
(15a) For the installed version of Python on a Mac there will be a Python folder in the Applications (Mac), double click the IDLE (highlighted in blue):


For the installed version on a PC the Python folder should be on the C drive (PC) in the Programs folder, double click the IDLE (highlighted in blue):

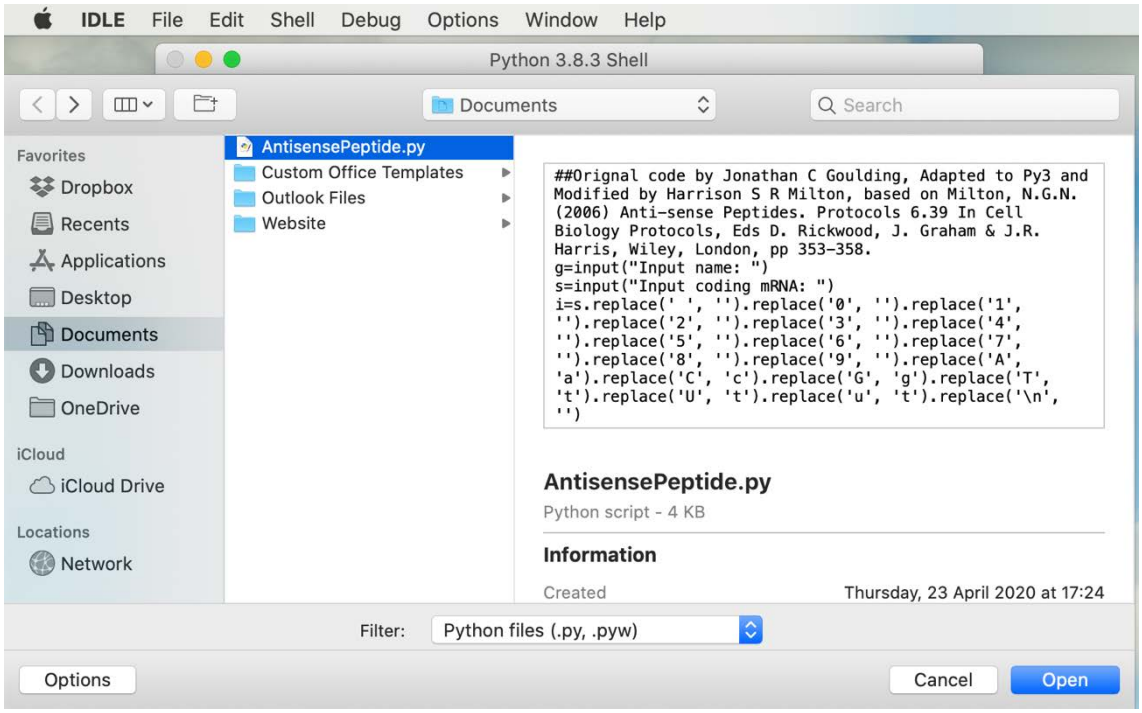



This will open a Python Shell and by clicking in File will give a pull down on which the Open (Ctrl+O) can be selected followed by selection of the AntisensePeptide.py file and then clicking open:

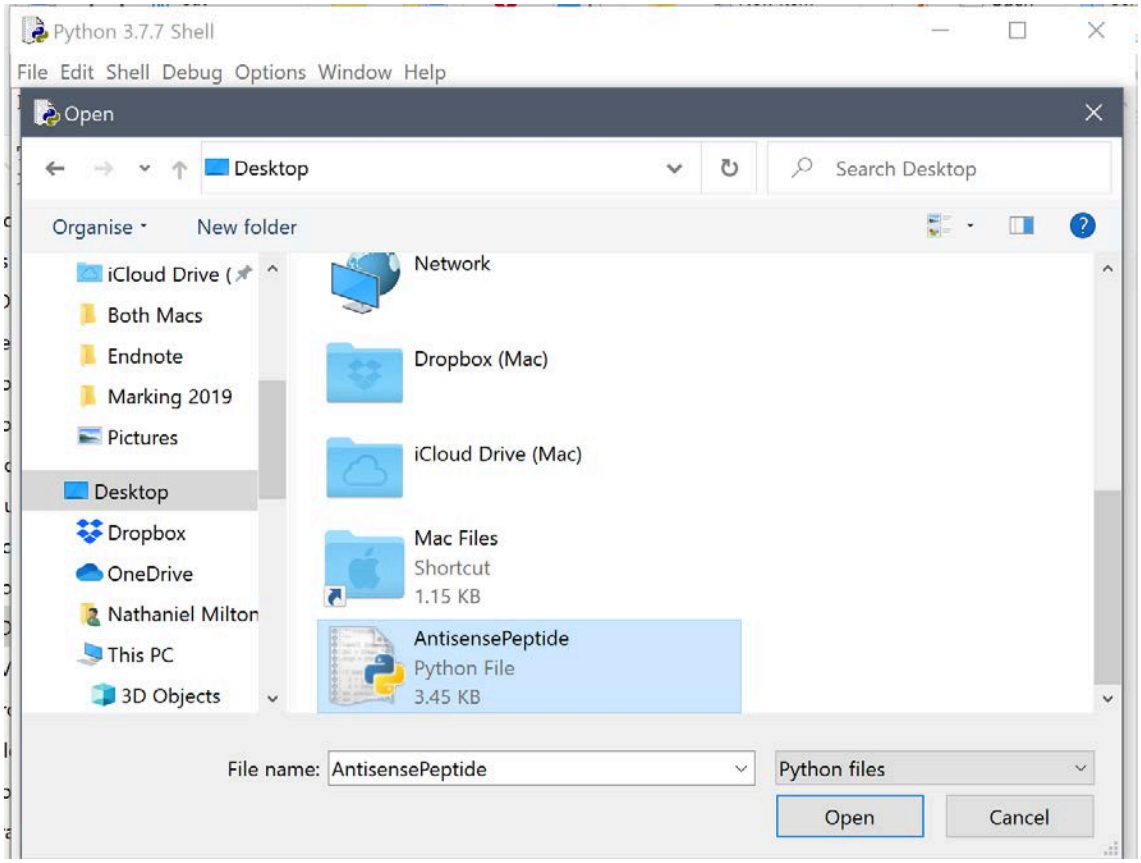


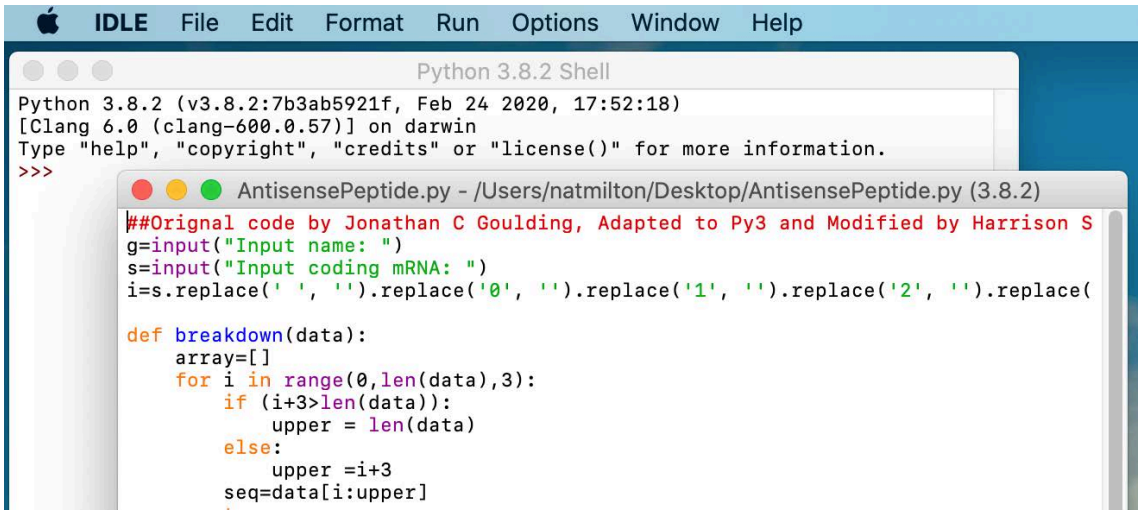
(i) Mac:


IDLE File Edit Shell Debug Options Window Help
Python 3.8.3 Shell
Documents
Search
Favorites
Dropbox
Recents
Applications
Desktop
Documents
Downloads
OneDrive
iCloud
iCloud Drive
Locations
Network
AntisensePeptide.py
Custom Office Templates
Outlook Files
Website
##Orignal code by Jonathan C Goulding, Adapted to Py3 and Modified by Harrison S R Milton, based on Milton, N.G.N. (2006) Anti-sense Peptides. Protocols 6.39 In Cell Biology Protocols, Eds D. Rickwood, J. Graham & J.R. Harris, Wiley, London, pp 353-358.
g=input("Input name: ")
s=input("Input coding mRNA: ")
i=s.replace(' ', '').replace('0', '').replace('1', '').replace('2', '').replace('3', '').replace('4', '').replace('5', '').replace('6', '').replace('7', '').replace('8', '').replace('9', '').replace('A', 'a').replace('C', 'c').replace('G', 'g').replace('T', 't').replace('U', 't').replace('u', 't').replace('\n', '')
AntisensePeptide.py
Python script - 4 KB
Information
Created
Thursday, 23 April 2020 at 17:24
Filter:
Python files (.py, .pyw)
Options
Cancel
Open

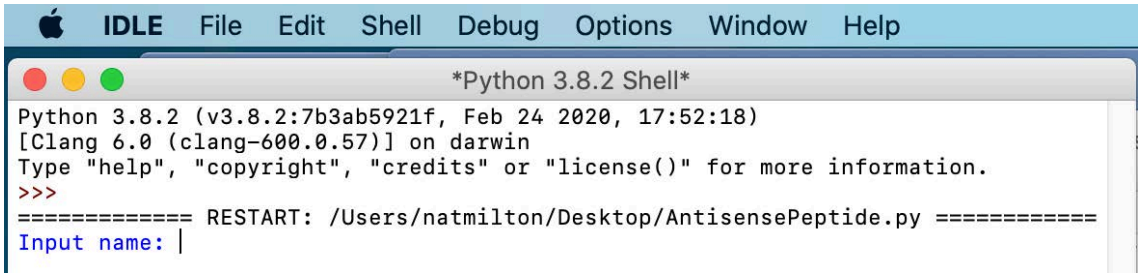
(ii) PC:


Python 3.7.7 Shell
File Edit Shell Debug Options Window Help
Open
Desktop
Search Desktop
Organise
New folder
iCloud Drive (
Both Macs
Endnote
Marking 2019
Pictures
Desktop
Dropbox
OneDrive
Nathaniel Milton
This PC
3D Objects
Network
Dropbox (Mac)
iCloud Drive (Mac)
Mac Files
Shortcut
1.15 KB
AntisensePeptide
Python File
3.45 KB
File name:
AntisensePeptide
Python files
Open
Cancel

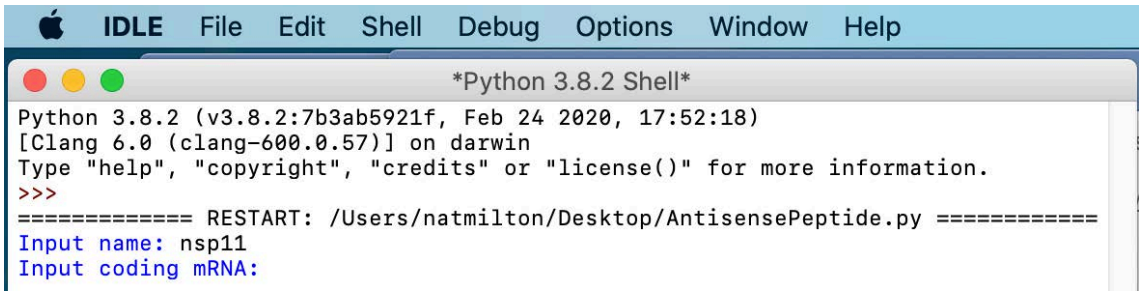
(15b) After opening AntisensePeptide.py a new window will appear, same for Mac and PC so from here onwards the protocol uses images from a Mac:




In the new window the Run function needs to be clicked and then the Run module F5 selected which will go to the following where the name on the target protein that the antisense peptides are generated against can be entered:

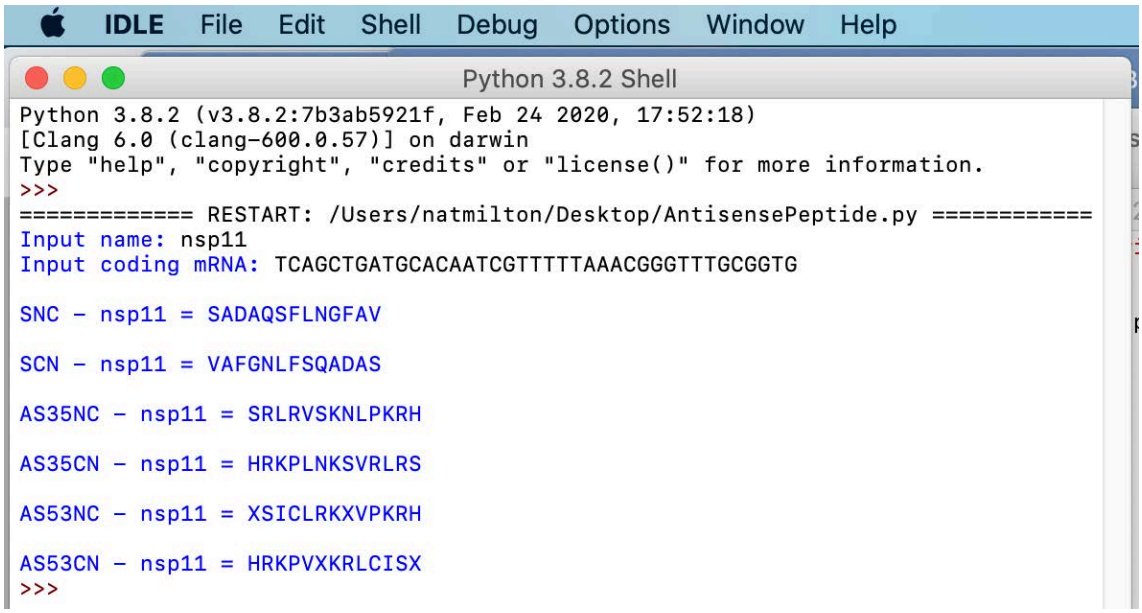



Once typed name hit return and will bring up following (nsp11 in this example):




The mRNA sequence used in the program needs to be in the format of a single word with no line breaks or paragraph marks ("¶") in the sequence (section 2d (page 10 above).

Then paste in mRNA sequence, for this example the sequence used is TCAGCTGATGCACAATCGTTTTTAAACGGGTTTGCGGTG, which is the nsp11 mRNA sequence and gives the following output:




In a very rare number of cases the a, t, c or g residues in the mRNA sequence could be replaced by an "n". This will cause an UNKNOWN to show in the peptide sequences which should be replaced by an X.

Where there is an * at the start (SCN, AS35CN and AS53CN) or end (SNC, AS35NC, AS53NC) of a sequence this is where the STOP codon was in the mRNA and can be deleted form the sequences used to run BLAST searches. If there is an * or an UNKNOWN in the middle of a sequence this indicates a problem with the mRNA used as these should only be at the end of coding sequences. Suggests a need to repeat section 2a-2c (see pages 5-10 above) to get the correct CDS mRNA component, particularly check that section 2c (see pages 8-10 above) to create an mRNA sequence that is a single word has been completed properly.

(15c) Copy the text from Input name down to the end of the AS53CN sequence and paste into a word document, save the Python outputs file with suitable name. These are the sequences that will be used for BLAST searches in sections 4 (see pages 14-16 above) and 5 (pages 17-19 above).

# 16: Appendix 3 - Manual antisense peptide generation

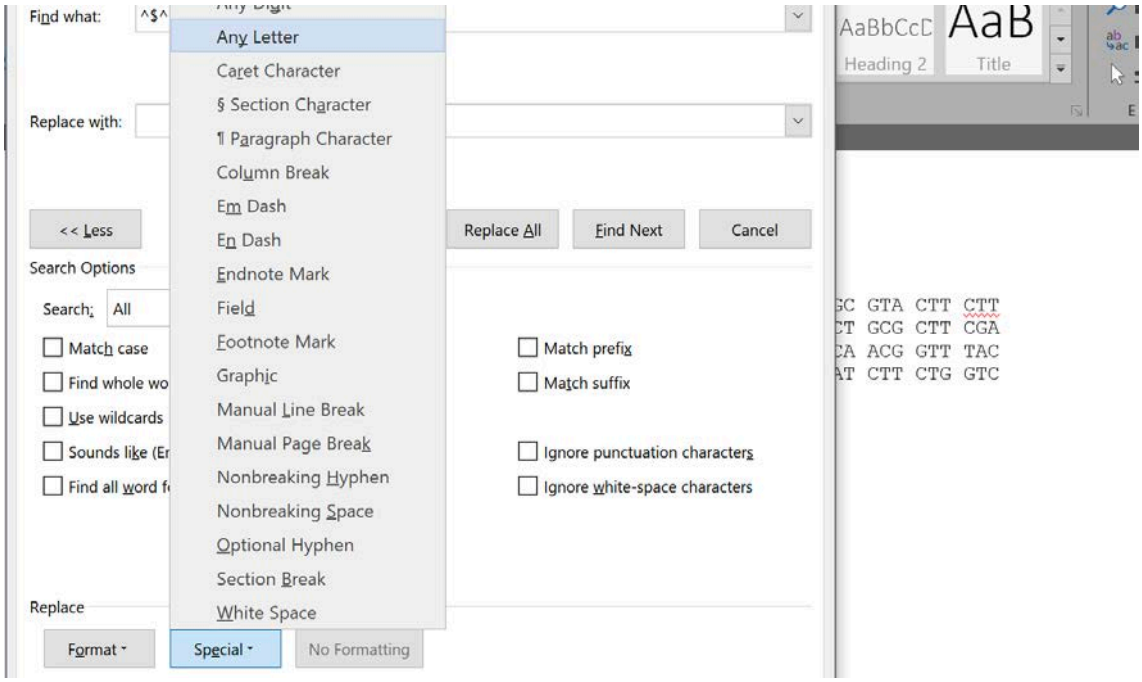
(16a) The AntisensePeptide.py when run online in Python (see section 3 pages 11-13) performs a series of tasks with the mRNA sequence. As an alternative method where Python cannot be installed or run online the antisense sequences can be generated manually (Milton 2006).

(16b) Initially, taking the mRNA sequence as a single string of text a space needs to be inserted between the 3rd and 4th base to convert the text into the coding triplet. This then needs to be repeated for the whole of the mRNA sequence. (in this example a short sequence peptide for nsp11 and its corresponding mRNA sequence has been used):

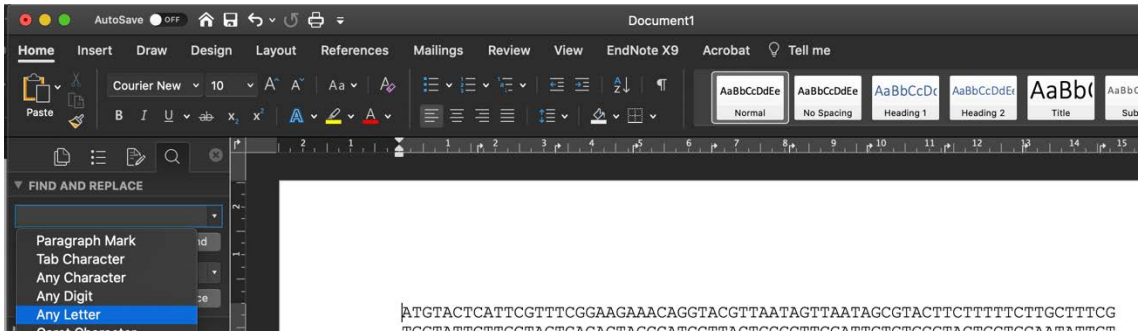
| | |
|---|---|
| mRNA: | tcagctgatgcacaatcgtttttaaacgggtttgcggtg |
| mRNA triplets: | tca gct gat gca caa tcg ttt tta aac ggg ttt gcg gtg |

(16c) To do this for a long sequence this can alternatively be carried out using find and replace. The cursor should be moved to the start of the first codon (often ATG) and then find and replace used as follows. Using the find and replace command in word as above select Any Letter, then repeat until three Any Letters are selected.

PC

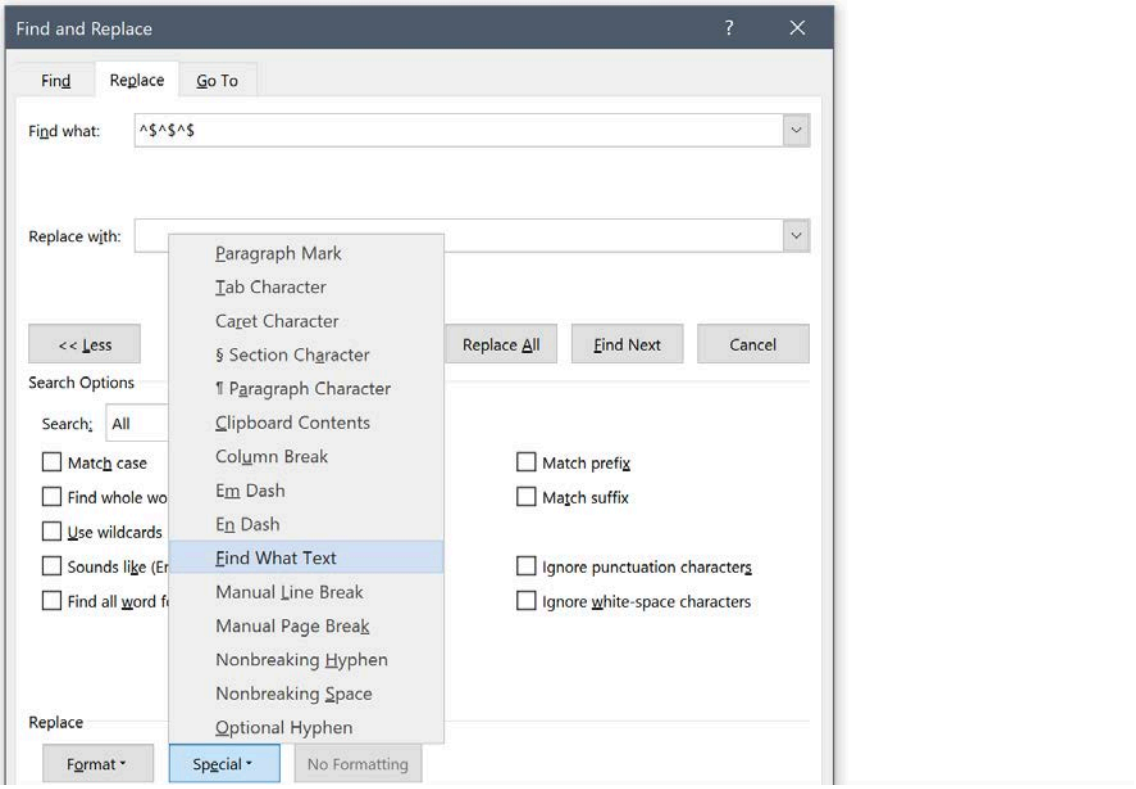


Mac

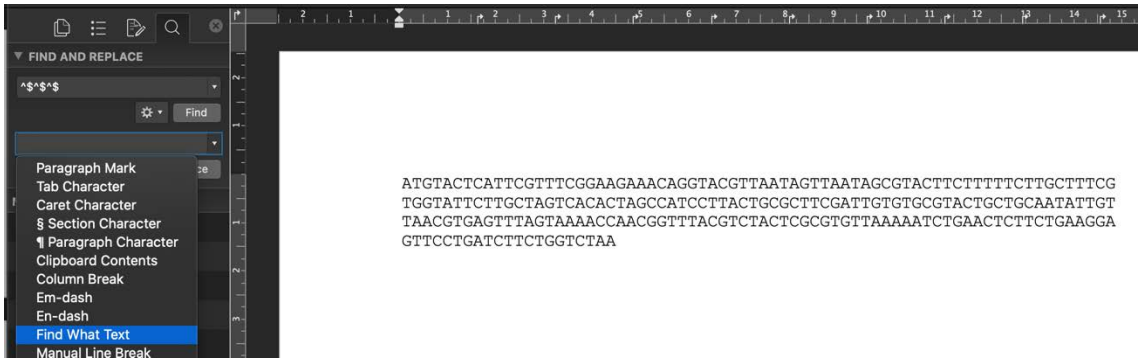



(16d) In the replace box use the Special tab and select "Find What Text" and then put a space after that.

PC



Mac




(16e) This will generate a sequence as follows, which should be saved and then a copy saved:

```
ATG TAC TCA TTC GTT TCG GAA GAA ACA GGT ACG TTA ATA GTT AAT AGC GTA CTT CTT
TTT CTT GCT TTC GTG GTA TTC TTG CTA GTC ACA CTA GCC ATC CTT ACT GCG CTT CGA
TTG TGT GCG TAC TGC TGC AAT ATT GTT AAC GTG AGT TTA GTA AAA CCA ACG GTT TAC
GTC TAC TCG CGT GTT AAA AAT CTG AAC TCT TCT GAA GGA GTT CCT GAT CTT CTG GTC
TAA¶
```

(16f) The sense coded target protein sequence should be available, however, if it needs to be generated from the mRNA this can be carried out using find and replace in word and replacing each triplet with the single letter code for the respective amino acids derived from the following tables for Sense strands:

| RNA triplet | Sense Amino Acid | RNA triplet | Sense Amino Acid | RNA triplet | Sense Amino Acid | RNA triplet | Sense Amino Acid |
|---|---|---|---|---|---|---|---|
| AAA | K | CAA | Q | GAA | E | TAA | * |
| AAC | N | CAC | H | GAC | D | TAC | Y |
| AAG | K | CAG | Q | GAG | E | TAG | * |
| AAT | N | CAT | H | GAT | D | TAT | Y |
| ACA | T | CCA | P | GCA | A | TCA | S |
| ACC | T | CCC | P | GCC | A | TCC | S |
| ACG | T | CCG | P | GCG | A | TCG | S |
| ACT | T | CCT | P | GCT | A | TCT | S |
| AGA | R | CGA | R | GGA | G | TGA | * |
| AGC | S | CGC | R | GGC | G | TGC | C |
| AGG | R | CGG | R | GGG | G | TGG | W |
| AGT | S | CGT | R | GGT | G | TGT | C |
| ATA | I | CTA | L | GTA | V | TTA | L |
| ATC | I | CTC | L | GTC | V | TTC | F |
| ATG | M | CTG | L | GTG | V | TTG | L |
| ATT | I | CTT | L | GTT | V | TTT | F |

The resultant sense sequence is the mRNA encoded peptide in the N-terminus to C-terminus direction (SNC). From the example of nsp11 above this would be:

SNC - nsp11 = SADAQSFLNGFAV

(16g) To generate the AS35NC antisense peptide sequence each triplet with the single letter code for the respective amino acids derived from the following tables for Antisense strand read in the 3'-5' for each amino acid:

| RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid |
|---|---|---|---|---|---|---|---|
| AAA | F | CAA | V | GAA | L | TAA | * |
| AAC | L | CAC | V | GAC | L | TAC | M |
| AAG | F | CAG | V | GAG | L | TAG | * |
| AAT | L | CAT | V | GAT | L | TAT | I |
| ACA | C | CCA | G | GCA | R | TCA | S |
| ACC | W | CCC | G | GCC | R | TCC | R |
| ACG | C | CCG | G | GCG | R | TCG | S |
| ACT | X | CCT | G | GCT | R | TCT | R |
| AGA | S | CGA | A | GGA | P | TGA | * |
| AGC | S | CGC | A | GGC | P | TGC | T |
| AGG | S | CGG | A | GGG | P | TGG | T |
| AGT | S | CGT | A | GGT | P | TGT | T |
| ATA | Y | CTA | D | GTA | H | TTA | N |
| ATC | X | CTC | E | GTC | Q | TTC | K |
| ATG | Y | CTG | D | GTG | H | TTG | N |
| ATT | X | CTT | E | GTT | Q | TTT | K |

The resultant output is the mRNA encoded Antisense 3'-5' peptide in the N-terminus to C-terminus direction (AS35NC). From the example of nsp11 above this would be:

AS35NC - nsp11 = SRLRVSKNLPKRH

(16h) To generate the AS53NC antisense peptide sequence each triplet with the single letter code for the respective amino acids derived from the following tables for Antisense strand read in the 5'-3' for each amino acid:

| RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid | RNA triplet | Antisense Amino Acid |
|---|---|---|---|---|---|---|---|
| AAA | F | CAA | L | GAA | F | TAA | * |
| AAC | V | CAC | V | GAC | V | TAC | V |
| AAG | L | CAG | L | GAG | L | TAG | * |
| AAT | I | CAT | M | GAT | I | TAT | I |
| ACA | C | CCA | W | GCA | C | TCA | X |
| ACC | G | CCC | G | GCC | G | TCC | G |
| ACG | R | CCG | R | GCG | R | TCG | R |
| ACT | S | CCT | R | GCT | S | TCT | R |
| AGA | S | CGA | S | GGA | S | TGA | * |
| AGC | A | CGC | A | GGC | A | TGC | A |
| AGG | P | CGG | P | GGG | P | TGG | P |
| AGT | T | CGT | T | GGT | T | TGT | T |
| ATA | Y | CTA | X | GTA | Y | TTA | X |
| ATC | D | CTC | E | GTC | D | TTC | E |
| ATG | H | CTG | Q | GTG | H | TTG | Q |
| ATT | N | CTT | K | GTT | N | TTT | K |

The resultant output is the mRNA encoded Antisense 5'-3' peptide in the N-terminus to C-terminus direction (AS53NC). From the example of nsp11 above this would be:

AS53NC - nsp11 = XSICLRKXVPKRH

(16i) The protein databases are always N-terminus to C-terminus orientation; however, proteins may interact with the binding site having one protein in the N-terminus to C-terminus orientation and the other in the C-terminus to N-terminus orientation. Hence the need to search the C-terminus to N-terminus orientation antisense peptides. To generate the SCN, AS35CN and AS53CN sequences they can be reversed using http://www.upsidedowntext.com/.


Using the sequence DSGYEVHHQKLVFFAEDVGSNKGAII as an example will generate an output of IIAGKNSGVDEAFFVLKQHHVEYGSD:

The overall output of sequences for the nsp 11 example should be:

SNC - nsp11 = SADAQSFLNGFAV

SCN - nsp11 = VAFGNLFSQADAS

AS35NC - nsp11 = SRLRVSKNLPKRH

AS35CN - nsp11 = HRKPLNKSVRLRS

AS53NC - nsp11 = XSICLRKXVPKRH

AS53CN - nsp11 = HRKPVXKRLCISX

# 17: Acknowledgements

The Antisense Peptide protocol has been constantly developed, initially from studies carried out by the author Dr Nat Milton during the original set up of NeuroDelta Ltd in 2003. The methods are based on his academic research carried out at Leeds Trinity University, Leeds Beckett University, the University of Westminster, the University of Roehampton, and University College London). The author would like to thank collaborators Dr Maria Ashioti, Tim Barnes, Dr Amrutha Chilimuri, Prof J. Robin Harris, Dr Farideh Javid, Dr Neema Mayor, Amanda Nercessian, Dr Mark Odell, Prof Jolanta Opacka-Juffry, Dr John Rawlinson, Dr Eridan Roche-Ferrera, Sabrina Werner, Prof Anthony F. Winder, and Dr Tamana Zemaryalai plus all of the other staff, students and collaborators who have supported this development and contributed to the research undertaken using these methods. I would also like to thank Jonny Goulding and Harrison Milton for the work on the Python Script that has speeded up the generation of antisense peptides dramatically plus Alexandra Fragkoulaki for the Excel sheets for Molecular Recognition.

The Antisense Peptide protocol methods have been tested by BSc, MSc and PhD students as part of their research projects plus taught Bioinformatics sessions at Leeds Beckett University, the University of Westminster, the University of Roehampton, and University College London. The author would like to thank all these students for their hard work and dedication during their studies, plus would like to congratulate them on their graduations and subsequent career progression. The author would particularly like to thank Dr Naghmeh Nikkheslat for encouraging the author to start running Bioinformatics undergraduate research projects and being the initial trainee on this programme whilst she was an undergraduate at the University of Roehampton.

The financial support by NeuroDelta Ltd and a U.K. Department of Trade and Industry Grant (LOT/0311684) that supported the initial work on the development and application of Antisense Peptides methods described in this protocol is gratefully acknowledged.