# Summary Antisense Peptide with Molecular Modelling Protocol

Background:

Computer based prediction of protein-protein interactions is a valuable *in silico* tool in a biological setting. Antisense peptide (or Complementary Peptides) sequences are derived from the complementary strand of DNA encoding a given protein, 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. The basis of antisense peptides is illustrated below:
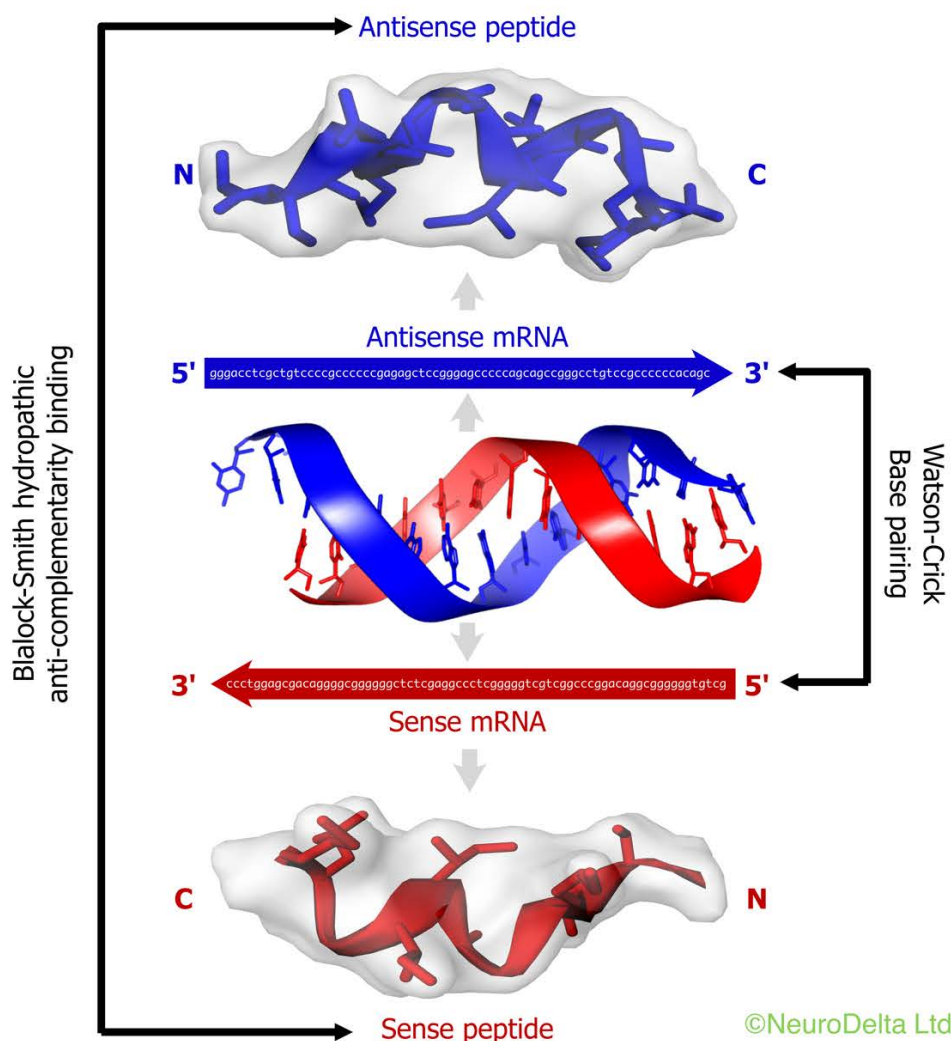
The Molecular Recognition Theory is based on a series of observations that protein sequences derived from the Sense Strand of DNA bind to protein sequences derived from the corresponding Antisense Strand of the DNA.

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. The antisense peptides have been shown to bind with high affinity to the given target protein due to hydropathic interactions, which is basically an interaction between water loving and water hating amino acids.

This protocol details the generation of antisense peptide sequences against a selected target and then using them to identify possible protein-protein interactions. The method uses a python script to generate the antisense peptides against the selected target based on the antisense peptides methods 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. Keys to confirming the validity of identified interactions are the *in vitro* or *in vivo* studies to determine that they actually occur, for example catalase binding to amyloid-ß and the role of this interaction in preventing the toxicity of amyloid-ß combined with demonstration in an Alzheimer's patient.

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. The antisense peptides themselves have been used as binding peptides to modify the actions of the target protein, for example ACTH. Antibodies raised against antisense peptide sequences have also been used to identify binding proteins *in vitro* or *in vivo*, for example LHRH.

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 method used for this has been described by and uses an algorithm to predict the protein-protein interactions with input of data from the identified interactions.

This is a summary version of the step by step Antisense Peptide protocol available from https://www.neurodelta.uk/resources/Full-Protocol-NGNM290724.pdf, for each section of this summary the full methods are referenced to the appropriate section of the full protocol. It should be used alongside the Antisense-Peptide.py file (https://www.neurodelta.uk/resources/BioinformaticsProtocolScript.py), run using a Python complier (https://trinket.io/python3), plus the ZDock server (http://zdock.umassmed.edu) for molecular modelling. The step by step Antisense Peptide protocol also contains links to a number of references relevant to the techniques described plus website links that contain useful information. A fully referenced version of the Background section above is also included.

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 or Mac OS up to version 11 (macOS Big Sur). The protocols have been written for use in Microsoft Word but can be adapted for other word processing software. The Python compiler for antisense peptide generation and protein docking software packages used are compatible with most browsers.

(i)   Generation of antisense peptides 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.

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

(v)   3D molecular modelling of the potential protein-protein interactions with the ZDock server using the potentially interacting residues identified in BLAST searches as a basis for the contacting residues in the model.

1   Target protein mRNA search (Full protocol section 2; pp 5-11):

(1a)   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 by searching the on the https://www.ncbi.nlm.nih.gov/nucleotide/ site. The ideal search will use the species name (for example Homo sapiens) plus the full protein name. If a box with a Gene title comes up look for the "RefSeq transcripts" link. If that doesn't show, select "mRNA" in the "Molecule types" side panel. The transcript of interest should be selected, generally select the longest complete transcript.

(1b)   Look for the CDS link within the document, this represents the region of mRNA that codes for protein and is the component of an mRNA sequence needed for antisense peptide generation. Clicking on the CDS will highlight the required sequence that can be copied. Clicking on the "FASTA" link on the lower RHS of the page will go to a new page with the mRNA sequence, starting at ATG. The whole of this sequence can be copied to a Microsoft Word file and used for antisense peptide generation. Make a note of the "NM_….." number plus the residue numbers and transcript details above the "ATG….." sequence as these will help find the information online if required.

(1c)   Using Microsoft Word the "¶" at the end of each line of sequence should be deleted.

(1d)   If there is a stop codon triplet at the end of the text, either TAA, TAG or TGA these three letters can be deleted as they are not specifically needed for antisense peptide generation. Then save file and use this version of the mRNA sequence for Antisense peptide generation.

2   Antisense peptide generation (Full protocol section 3; pp 12-14):

(2a)   Stage two uses the Antisense-Peptide.py file (https://www.neurodelta.uk/resources/BioinformaticsProtocolScript.py) run in an online Python compiler (https://trinket.io/python3) to generate antisense peptide sequences. This can be run in most browsers on a Mac, PC, Chromebook, iPad etc and has been tested using both Safari and Google Chrome.

(2b)   The text of the Antisense-Peptide.py file (https://www.neurodelta.uk/resources/BioinformaticsProtocolScript.py) should be copied and pasted into the compiler.

(2c)   The compiler can then be run using the run function which will bring up the "Input Name:" command, 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 1d above can then be pasted in and the program run.

(2d)   Copy the text from Input name down to the end of the AS53CN sequence and paste into a word document and save the Python outputs word file with a suitable name. These are the sequences that will be used for BLAST searches below.

3   Antisense peptide BLAST searches (Full protocol section 5; pp 18-20)

(3a)   In stage three separate BLAST searches should be carried out for each antisense sequence:

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

Using the https://blast.ncbi.nlm.nih.gov/Blast.cgi website 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)".

Next click on "Algorithm parameters" and 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. Then click on BLAST and wait for results to appear.

(3b)   After the BLAST search has finished the screen will an "RID" tab just below the Job title. Click on the Download All dropdown and select Text. Save txt file to an appropriate folder (this file will be used for analysis). A screenshot or save as pdfs file can also 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.

4 BLAST data extraction (Full protocol section 6; pp 21-31)

(4a) 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.

(4b) 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.

(4c) Within the txt file data downloaded from (3b) above the key elements should be extracted into a table with the following headings:

Name of protein similar to antisense peptide	Protein ID	Size	Residues	Query residues	% ID	% +ve	% Gaps

Using the sections highlighted in red below will produce the table that follows:

```
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
```

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

(4d) Make sure that none of the proteins have a % Gaps > 0 as this data should be discarded.

(4e) If the antisense peptide sequence screened is an AS35NC or AS53NC the Query residue numbers correspond directly to the SNC residue numbering. However, in this case an AS35CN or AS53CN peptide the Query numbers will need to be converted as described in 4b above to give a table like the following example:

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

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

(4f) Within the results files there are other potentially useful pieces of information that can be used.

5 Molecular recognition analysis (Full protocol section 7; pp 32-40)

(5a) 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

(5b)   The Molecular Recognition (MR) scoring uses the above table (section 5a page 7) as 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.

(5c)   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 (from 3b above) 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 5b above 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 MR score is best expressed as a % of the total number of residues in the target sequence. Within the results table (see 4e, above) 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

6   PDB files for ZDock modelling (Full protocol section 8; pp 41-52):

(6a)   For molecular modelling protein structure files in a PDB format are used. A Protein BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) can be run using the protein ID for the target or interacting protein for the "Enter Query Sequence" and searching in the Protein Data Bank proteins (pdb) "Database" in order to find similar structures. Using the "Alignments" tab in the results it is possible to find structures that cover regions of interest.

(6b)   PDB files can be obtained from a number of sources, the RSCB Protein Databank (https://www.rcsb.org) is a preferred choice and contains validated structural

information, often with references. 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).

(6c)   If 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).

(6d)   For protein sequences where no PDB file is available a predicted structure can be created using I-Tasser (https://zhanglab.ccmb.med.umich.edu/I-TASSER/). A sequence between 10 and 1500 amino acids can be used for this purpose.

(6e)   It is important to review the structure PDB files to ensure they cover the sequences of the given proteins identified in the BLAST searches, also to confirm that the residue numbering matches the protein sequence numbering used in the BLAST search results. Many models are based on post-translationally modified proteins and may therefore lack regions cleaved as part of this process.

(6f)   For many models the PDB files contain more than one protein chain and may contain more than one protein. As part of the modelling, it may however be preferable to create a PDB file with only a single chain within it (or even part of a chain). This can be carried out using the iCnD program (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html).

(6g)   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.

To overcome these problems of PDB files containing the same chain labels one of the PDB files can have the chain name altered using the renaming chains website (http://www.canoz.com/sdh/renamepdbchain.pl). After uploading a PDB file with a single chain it is possible to change the chain label to any other letter not used in the other PDB, allowing for effective ZDock analysis.

7   ZDock interaction modelling (Full protocol section 9; pp 53-56):

(7a)   Launching the ZDock protein docking program (http://zdock.umassmed.edu) goes to a window where two PDB Files can be uploaded. An email address should be entered, 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 next to the Input Protein 1 or 2 is used to select PDB files that can be uploaded.

Once both files have been uploaded click Submit. The next stage is to select the residues that are to be part of the binding site. These residues are selected

under "Select Binding Site Residues" on the ZDock server. These are the residues from the BLAST search results file for the target protein and the identified interacting protein that is similar to the antisense peptide. 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.

(7b)   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. 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.

8   iCn3D interaction data (Full protocol section 10; pp 57-68):

The top 10 prediction PDB files from the ZDock analysis can also be viewed and analysed using the iCn3D protein structure viewer (https://www.ncbi.nlm.nih.gov/Structure/icn3d/full.html). Uploaded PDB files can be analysed using the "Analysis" tab followed by either the "View Sequences & Annotations" or "H-Bonds & Interactions" dropdown features to identify interacting residues and download full details of those.

9   EzMol Interaction images (Full protocol section 11; pp 69-73):

Image files can be created by uploading top 10 prediction PDB files from the ZDock analysis using the EzMol structure display (http://www.sbg.bio.ic.ac.uk/ezmol/). Uploaded PDB files can be manipulated to highlight specific components, including coloured or labelled residues plus surface structure display.

10   Interpretation of results (Full protocol section 12; pp 74-75):

(10a)   The target protein orininally chosen will have a protein id 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 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).

(10b) These key features also need to be identified for each of potential binding proteins listed in the results tables (see 5c above) from the BLAST searches for alignments with AS35NC, AS35CN, AS53NC and AS53CN sequences.

(10c) 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.

(10d) 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.

(10e) 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).

(10f) From the iCn3D protein-protein interaction data extraction (8 above) 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 5c 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.

(10g) Images of the suggested interacting structures from sections 8 and 9 (above) can be generated to illustrate the suggested interactions.

(10h) 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.

(10i) The antisense peptide sequences can also be used to generate synthetic peptides for use in experimental settings.

11   Acknowledgements