5BBB0226 Principles of Bioinformatics
Coursework Assignment 2 assigned date 28th March 2021
Prof Franca Fraternali
e-mail:franca.fraternali@kcl.ac.uk

Your King's ID:	**K20041606**

Deadline date: 28th April 2022 4 pm London time

Comparative Modelling and functional investigation of a protein with unknown structure.

This course work involves the analysis of structure and function properties derived from a given protein sequence (the UniProtKB/TrEMBL accession number or the FASTA file with the sequence will be given). Each student is allocated a different protein. You will write a guided essay describing the steps and results of homology modelling, as well as other bioinformatics investigations surrounding your assigned protein.

Login on KEATS. Click on the module title to open its page and you will find that you will have been assigned a protein for your coursework as well as this document.

Each student should complete the task alone.

Below we have listed the steps involved in this assignment. Here we guide your essay writing by listing the **tasks** (in *italics*) and **points to address** (in blue) for this assignment, with a suggestion of word count for each section. ***The word count does not include figure/table legends.***

Please write your answers in paragraph format and put answers in the boxes provided. NOTE: this assignment is an essay, NOT a 'lab report' documenting evidence of completing your work. Please explain and discuss your finding as appropriate.

Provide citations wherever necessary (and complete the bibliography section towards the end of this document).

Provide essential figures and tables throughout your answer wherever you deem suitable. You may put them inside or outside of the provided boxes. Make sure any texts or labels on your figures/tables are legible. Remember to include legends to explain these materials.

You may include extra materials (e.g. multiple sequence alignments) as appendix to this coursework. Attach them at the back of this document.

1) *Identification of the protein name and organism. If you are given a UniProt identifier, retrieve the sequence in FASTA format; if you are given a FASTA, retrieve the corresponding UniProt accession. This is your query sequence.*

a) Paste in the box below both the UniProt accession and the FASTA sequence of your assigned query. If your assigned sequence encodes only part of a protein, include only the sequence assigned to you, and specify which part of the UniProt entry with which it matches.
Outline the steps you have taken to retrieve these entries and information.

UniProt Accession
P40313

FASTA Sequence
>sp|P40313|CTRL_HUMAN Chymotrypsin-like protease CTRL-1 OS=Homo sapiens OX=9606 GN=CTRL PE=1 SV=1
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFC
GGSLISQSWVVTAAHCNVSPGRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNN
DVTLLKLASPAQYTTRISPVCLASSNEALTEGLTCVTTGWGRLSGVGNVTPAHLQQVALP
LVTVNQCRQYWGSSITDSMICAGGAGASSCQGDSGGPLVCQKGNTWVLIGIVSWGTKNCN
VRAPAVYTRVSKFSTWINQVIAYN

The protein assigned for this essay is called P40313 and it is a chymotrypsin-like protease. The query sequence of chymotrypsin-like protease from the gene CTRL on chromosome 16 in homo-sapiens is above. A UniProt identifier was given (P40313) and the FASTA file of the sequence was obtained (above). This will be the target protein for homology modelling. The FASTA sequence was retrieved from Uniprot.org by pasting the identifier code given into the search bar and formatting the protein into a canonical FASTA file through the website. The protein sequence assigned and retrieved codes for the entire chymotrypsin-like protease protein.

(100 words)

b) Using information from the UniProt entry, describe your assigned protein.

P40313 (CTRL_HUMAN entry name) is from the gene CTRL-1 located on chromosome 16 in homo-sapiens. It functions as a hydrolase, protease and serine protease (endopeptidase), secreted in the extracellular space during protein catabolism and proteolysis. It's primarily expressed in the pancreas but also 118 other tissues. It is a highly studied protein, and the alpha fold residue confidence score (pLDDT) is strong. The protein is 28,002 Da and 264 amino acids long: the signal peptide from 1-18, pro-peptide activation from 19-33 and the chymotrypsin-like protease domain from 34 – 264. This chain is the active component consisting of 3 key charge relay active site positions at amino acid 75, 121 and 214. It requires no cofactors for function. There are 2 potential isoforms which are 190 amino acids long and two missense variations in amino acids 150 and 173. There are some associated diseases in the mitochondrion and cell membrane. The human CTRL gene that encodes P40313 has binary interactions with 3 proteins: CLDN5, FAM209A, FATE1.

(max. 150 words)

2) *Perform homology modelling via web server: SWISS-MODEL.*

Describe what you have done for this step. Discuss the table of templates SWISS-MODEL returns. Identify any promising templates to be taken further to perform modelling of your target, and explain your selection.
(Hint: select at least 1 template here for automated homology modelling.)

SWISS-MODEL is an automated homology modelling technique where the FASTA of the target protein was input and the software provided 50 templates with accurate, experimentally determined structures from various PDB databases (including BLAST) to produce/predict a model. High template query coverage (80+ in this case) is ideal so coordinates can be safely passed between the 3D structure to the target for a large portion of the sequence. There also must be an identity above 30% to assume a shared common ancestor for the template to be a suitable match to the target.
Many templates fit these broad criteria which were narrowed down for a high resolution that can accurately portray side chains with confidence of atom locations (~2Å ideal) and high QMEAN scores (~80 in this case) which suggest low model residue deviation and high model accuracy. The selection came to two crystallised *bos taurus* templates:
2cga.1.A Chymotrypsinogen A (53.88 identity, 1.8Å resolution, homo-dimer, 0.79 GMQE) and *1p2m.3.C* Chymotrypsinogen A (53.88 identity, 1.75Å resolution, hetero-tetramer, 0.81 GMQE). The *2cga.1.A* template's additional 0.32 QSQE score (expected accuracy of modelling it as an oligomeric form) is irrelevant as only a value above 0.7 can be considered beneficial in following the predicted quaternary structure by estimation and our protein's oligomeric form is strictly predicted as a monomer. The lower GMQE, higher resolution and better QMEAN scores (-23 compared to -0.62) make *1p2m.3.C* a statistically superior model; it also has 6 sulphate ligands (**fig1**) but these are 'not biologically relevant' as the target protein requires no cofactors for its physiological function.

(max. 250 words)

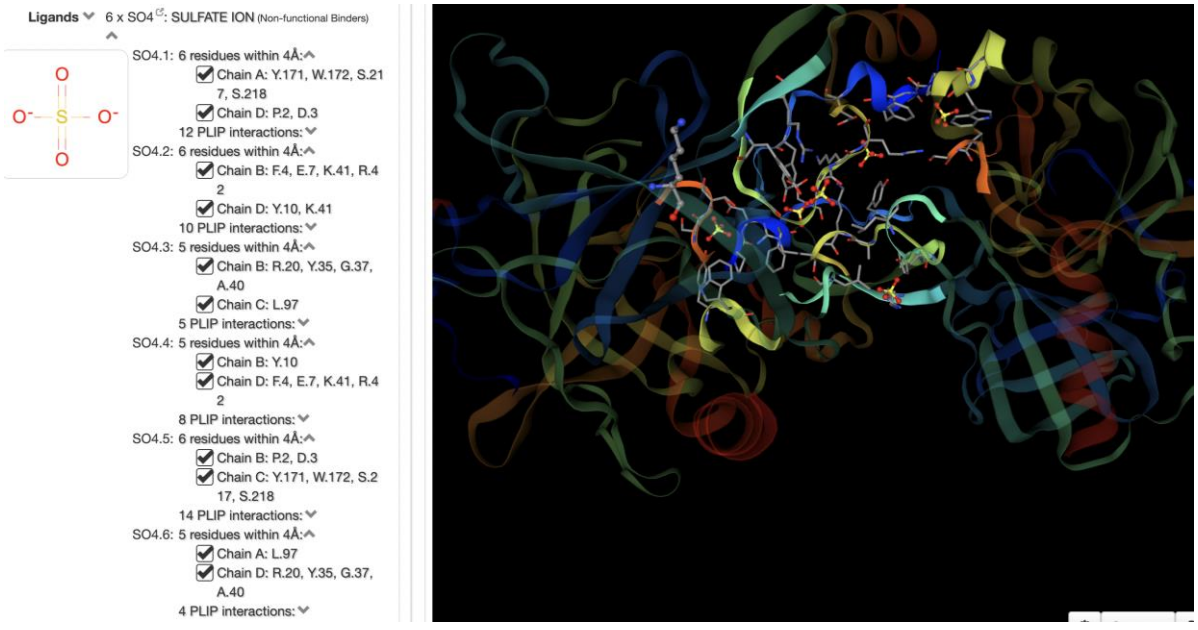


Figure 1: The 3D structure of *1p2m.3.C* Chymotrypsinogen A with ligands**.** This is a graphical image of the ligands in *1p2m.3.C* template from SWISS-MODEL. It visualises the 6 sulphate ligands within the 3D structure: these are removed in the modelling process. SWISS-MODEL confirms the 6 ligands as 'not biologically relevant'.

3) *Examine the model(s) created by the server. Download the PDB file(s) for the template(s) used by the server, as well as the PDB file(s) of the model(s) created by the server.*
(**Save these results, as well as any evaluation plots/statistics SWISS-MODEL returns for your model – you will need them later in this assignment.**)

4) *Perform a BLAST search with the target sequence for structure modelling (using PDB database of protein structures). Retrieve sequences of suitable templates.*

Describe the BLAST results.

A BLASTP of the target FASTA was performed against the PDB database, part of the alignment homology modelling procedure. The goal is to find suitable templates based on a series of factors; coverage/sequence identity which assess alignment quality; e-values/alignment lengths which assess score; experimental techniques/resolution which affect template quality; and other factors like protein type, ligands and missing data. The BLASTP output 100 sequences, which had query covers from 84-93%, E values from e^{-30} to e^{-72} and identities from 36-56% (**fig2.A**). This output has less variety than SWISS-MODEL, which uses templates from many databases including BLAST, like HHblits.

Difference between e values was mainly negligible as they were all low with significant hits. The query is 264 amino acids long, templates ranged from 128 to 791 amino acids and there was a large divide between alignment scores with two at 262 and 244 and the rest 160 or below (**fig2.B**, red). The two templates with the highest identity to the target sequence had low query covers of 48% so these aren't considered; query coverage is essential to produce a model with accurate template coordinate mapping. The sequences with best identities (40%+) did not come from homo-sapiens and instead came from other species (**fig2.A,** scientific name). Bovine Chymotrypsin A sequences had the best balance between identity, query coverage and e-values; likely due to their close evolutionary relationship to the target via the S1A peptidase domain [*INTERPRO*, 2022]. There are some homo-sapiens sequences with a very high coverage of 93% but they have a lower identity (max 37%) and they're plasminogen proteins; chymotrypsin-like proteins are far more relevant.

(max. 250 words)

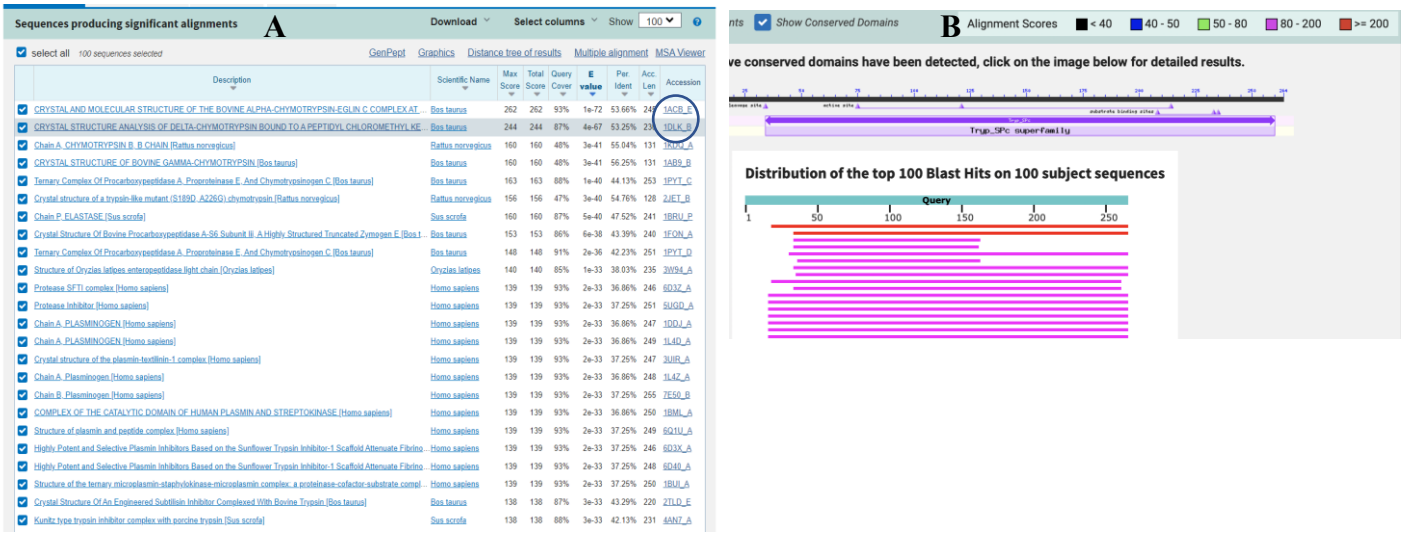


Figure 2: The NCBI BLAST results against the PDB database A: This is an image of the PDB templates produced by BLASTP. It ranks the templates by e value (smaller values at the top) and shows the first 22 templates of 100. **B:** This is the Graphic Summary. It visually ranks the alignment scores of the templates: the templates in red with the best scores (greater than 200) are most suitable for homology modelling.

5) *From your BLAST search, select the best template(s) for homology modelling.*

(Hint: select at least 1 template from the BLAST table.)

Describe which template(s) you have selected, and justify your selection.

The e value of $1e^{-72}$ for *1ACB_E crystal bovine alpha-chymotrypsin-eglin c complex* stands out as the lowest. Furthermore, the coverage is high at 93%, making coordinate mapping to the template accurate, and it's ranked 4th in identity at 53.66% which comfortably passes the 30% threshold for homology modelling. It's 245 amino acids long so it is a suitable match for the 264 amino acid target sequence, and it also has a high pairwise alignment score (**fig4**). The closest competitor would be *1DLK_B delta-chymotrypsin bound to a peptidyl-chloromethyl-ketone* (Cl^-) template of 230 amino acids, missing 34 amino acids from the beginning (including the signal peptide and activator domains). It has an identity of 53.25, slightly lower coverage of 87% and a slightly lower e value of e^{-67}. The Cl^- ligand may add some uncertainty to the homologous target protein structure which has no cofactors.
Both sequences stand out as the highest in length and score (**fig2.B**). They were also resolved by X-RAY crystallography methods preferred to NMR and cryo-EM. The protein data bank (**fig.3**) can further discriminate by resolution: *1DLK_B* has a worse resolution of 2.14Å compared to *1ACB_E* (2Å). The better resolution means *1ACB_E* can more accurately portray side chains for confidence of atom locations in structure. Overall, *1ACB_E* is statistically more suitable and should be favoured; it also has no ligands and which would provide a good comparison to the sulphate bound *1p2m.3.C* template.

(max. 200 words)

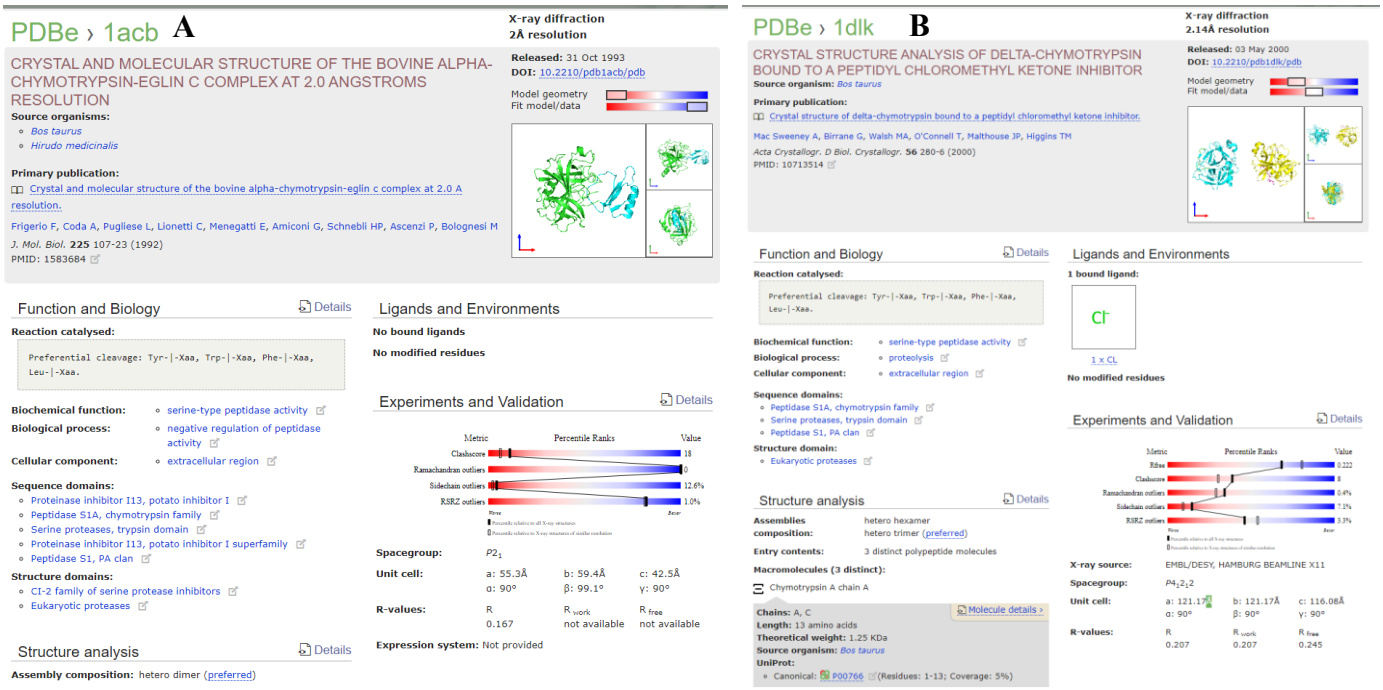


Figure 3. The two best templates were hard to separate by BLAST parameters so further details such as resolution were obtained from the Protein Data Bank Europe on EBI: A: This is the full PDB data of the template 1ACB_E from BLAST. It shows the resolution, ligands, function, domains, and experiments related to the protein and its model. **B:** This is the full PDB data from the template of 1DLK_B which portrays the same data.

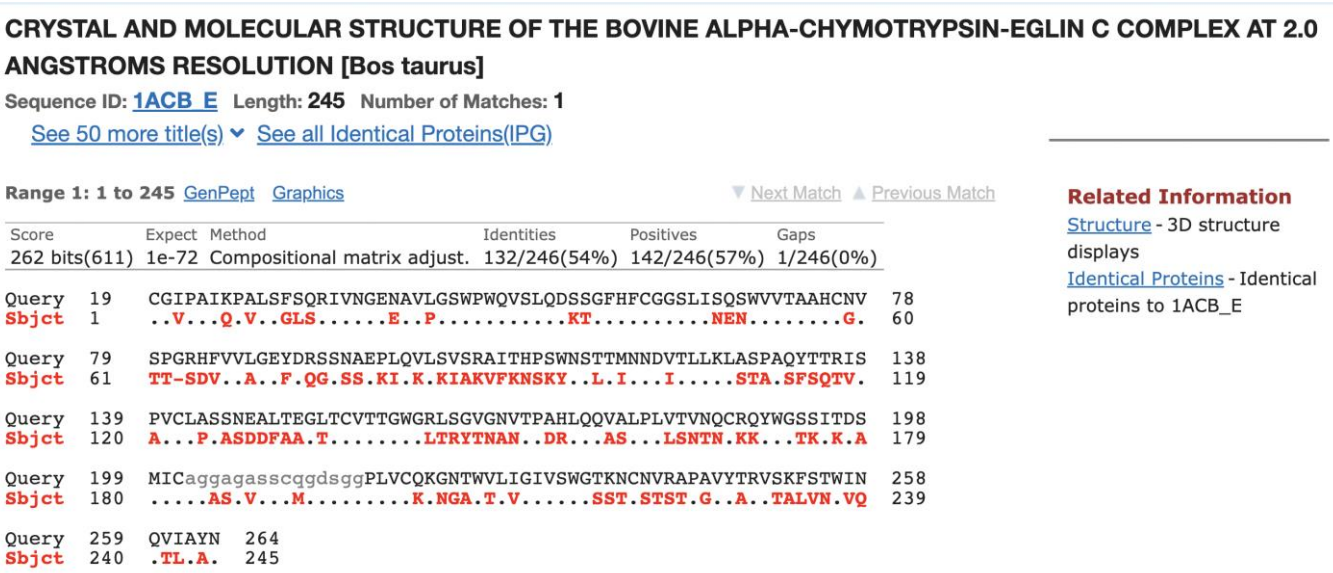


Figure 4. Pairwise multiple sequence alignment with dots for identities: This diagram shows mismatched residues as red and matched residues as dots. It visualises the alignment for the statistically superior and chosen template *1ACB_E.*

6) Perform T-Coffee alignments of your query sequence and the selected templates. Compare the template(s) hits returned from SWISS-MODEL in step (2), and the one(s) you would have selected based on BLAST and the T-Coffee alignments. Are there any notable differences between T-coffee alignments and SWISS-MODEL alignments?

T-Coffee uses sequence-template approaches and a consistency measure: alignments between the *1ACB_E/1DLK_B* BLAST templates and the query sequence produced mutual scores of 993: a 99.3% consistency with the primary library. The fasta_aln files were excellent aside from 2 and 1 'average' regions in *1DLK_B* and *1ACB_E* respectively (**fig6**). A specific signal peptide domain from the target homo-sapiens FASTA (amino acids 1-18) was missing in both templates due to an N-terminal truncation (**fig6,** black arrow): the templates are different species to the target with different signalling mechanisms. *1DLK_B* was missing an additional 15 amino acids in this truncation which included the activation domain of the target sequence (amino acids 19-34), making it less appropriate for homology modelling (fig6, red circled). T-Coffee ranks alignment via the *Gonnet PAM 250 matrix:* asterix (conserved residue), period (mismatch with weakly similar properties) and colon (mismatch with strongly similar properties).

SWISS-PLOT results are far more representative of the 3D PDB structure than T-COFFEE. The N-terminal truncation of *1p2m.3.c* (**fig5.A**) is the same as *1ACB_E* as it is also missing the 1-18 signal peptide. SWISS-PLOT identified the first 15 amino acids of the *1p2m.3.C* sequence (truncated equivalent to the target's 19-34 amino acid pro-peptide activation domain) as a pancreatic trypsin inhibitor. The structural-template alignment in SWISS-PLOT of *1ACB_E* (**fig5.B,** via its T-COFFEE fasta_aln) instead had an Eglin-C serine protease inhibitor domain; of their mutual 2 small gaps, the first gap in *1ACB_E* pro-peptide activation domain is three amino acids shorter (**fig5**, circled) and this difference in alignment changes the domain in modelling. The target protein is specifically a zymogen endopeptidase that is released in the pancreas for digestive function. As a result, the pancreatic pro-peptidase activation domain template *1p2m.3.c* is far more biologically relevant to target and would provide a more suitable model prediction (**fig14**) [*RefSeq*, 2016].

Statistically, the *1ACB_E* alignment has the same coverage as *1p2m.3.C* at 93%, sequence similarity at a sufficient 0.47 and identity at 53.88. *1pm2m.3.C* and *1ACB_E* are also both 245 AA long and conserved active site residues are present in both models at key amino acid positions (H at 75, D at 121 and S at 214). The template sequences are identical so the difference in SWISS-MODEL structural-template alignment is likely due to the 6 sulphate ligands that can alter the predicted 3D structure modelled from *1p2m.3.c* (**fig1**), regardless of their lack of 'biological relevance' to the endopeptidase function of the target which doesn't require cofactors. As expected, the sequence alignment via BLAST alignment homology modelling (*1ACB_E*) is generally better than SWISS-MODEL'S with smaller gaps but the *1pm2m.3.C* resulting model is more biologically relevant.

(max. 400 words)

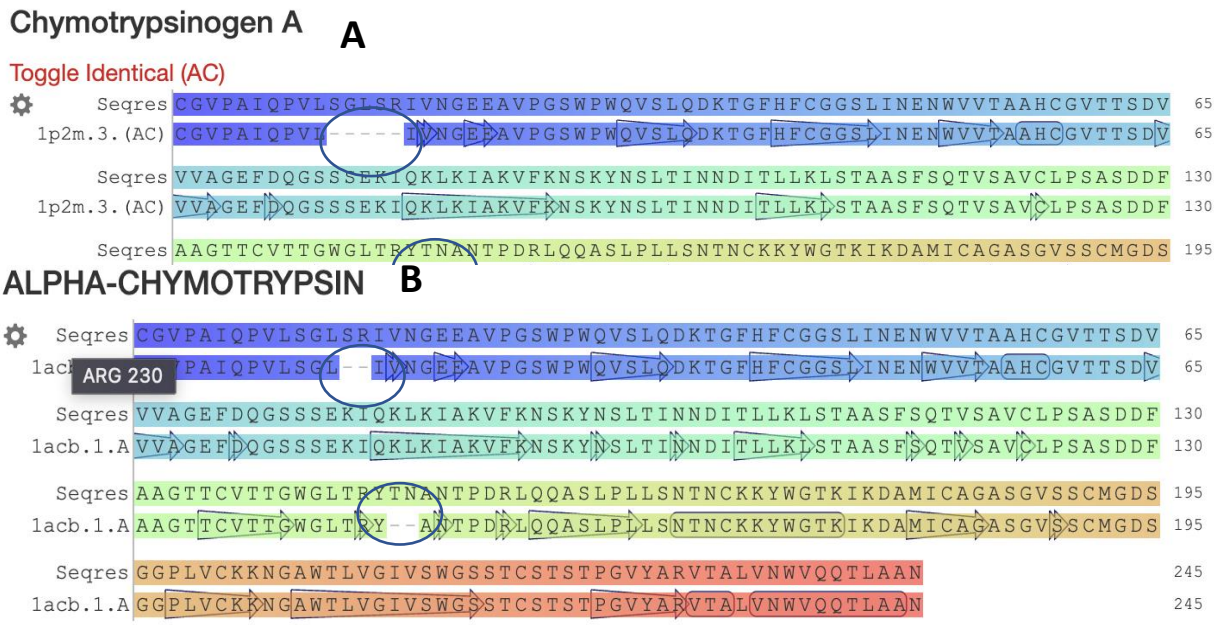


Figure 5. A: The model-template alignment of *1p2m.3.c* from SWISS-MODEL. Gaps are circled: a 5 and 2 amino acid gaps. **B:** The model-template alignment of *1ACB_E* from BLAST on SWISS-MODEL. Gaps are circled: 2 and 2 amino acid gaps.

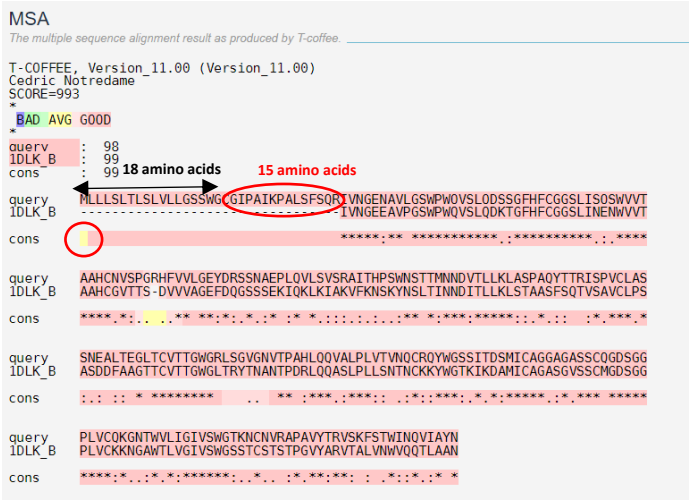


Figure 6. This shows the T-COFFEE alignment of the *1DLK_B* template: It is identical to 1ACB_E other than key additional negative features circled in red: the notable missing pro-peptide domain and the additional region of average alignment. The missing signal peptide region of both sequence is displayed with a double-sided arrow.

7) *Perform your search of a 3D model via Alphafold2- extract the domain you are interested in*
(**Save these results, as well as any evaluation plots/statistics SWISS-MODEL returns for your model – you will need them later in this assignment.**
8) *Examine your models from homology modelling and Alpafold2 and the template structures using molecular visualisation software (e.g. PyMOL). Perform a comparison (structural, 3D superimposition) with the corresponding templates, and/or any deposited models (e.g. in databases such as Modbase) of your target protein.*

a) Display informative snapshots of your structural analysis. Discuss these results, referring, wherever necessary, to other results you have had so far, e.g. the alignments and the BLAST/SWISS-MODEL template hits table.

The target domain of Alphafold2 P40313 for extraction is amino acids 19-264 as this is the conserved region – ALPHAFOLD2 covers the entire *UniProt* sequence and the lack of the signal peptide domain in experimental structures closely related to the homo sapiens target sequence mean this region is very low in pLDDT (confidence score for a predicted structure) and is a poor yellow/orange colour (**fig.7A**). Otherwise (amino-acid 19 onwards), the ALPHAFOLD has a very good pLDDT and a dark green PAE suggesting little aligned error (**fig7.B**).

The SWISS-MODELs of P40313 chymotrypsin generated from the experimentally determined templates can be superimposed as they have a decent sequence similarity, >30%. The models produced from *1p2m.3.C* and *1ACB_E* were high resolution (the templates are 1.75Å and 2.00Å respectively) and thus so was the superimposed structure (**fig8.C**). Qualitatively, alignment was very good for the b-sheets and alpha-helices in the structure and the chains merged in colour: they are both bovine Chymotrypsin-A templates (**fig9.B**). The active site is also evolutionary conserved between chymotrypsin proteins: the 3 key resides in the charge-relay system at positions 75, 121 and 214 are closely overlapping in stick form (**fig.9.C**). Homology modelling preserves the secondary and tertiary structure more when predicting structure as it is not favourable to open gaps in motifs; instead, they are in loop structures that are less relevant to function, evident from **fig8** which has far more misalignment in loops (black circles) than alpha-helices and beta-sheets (red circles). However, even they were well aligned between the template derived models (**fig9.A**) with a low overall RMSD (Root-Mean-Square-Deviation) of 0.287.

The extracted ALPHAFOLD domain (**fig7.C**) was compared to the *1p2m.3.C* and *1ACB_E* models produced by SWISS-PLOT of the same amino acid region. The structures were aligned and then super-imposed which produces a RMSD. The ALPHAFOLD2 is not an experimentally determined structure so its resolution in *Pymol* is much lower than the crystallised structure templates obtained from SWISS- MODEL. When visually comparing the 3D models predicted from the experimentally determined templates generated by SWISS-MODEL to ALPHAFOLD, there was far more disparity between the structures than template-template model comparisons: visibly more misaligned looping circled in **fig8.A** and **fig8.B** than **fig8.C**. The RMSD for automated homology modelling (*1p2m.3.C*) was closer to ALPHAFOLD than *1ACB_E*, at 0.478 as opposed to 0.514. Since this ALPHAFOLD structure is made to be 100% homologous to the actual homo-sapiens target sequence, this suggests that the *1p2m.3.C* model is biologically more like the target. This confirms previous observations of better biological similarity of the automated homology modelling *1p2m.3.C* domains like the pancreatic trypsin inhibitor which aligns with the pancreatic digestive function of the target.

(max. 400 words)

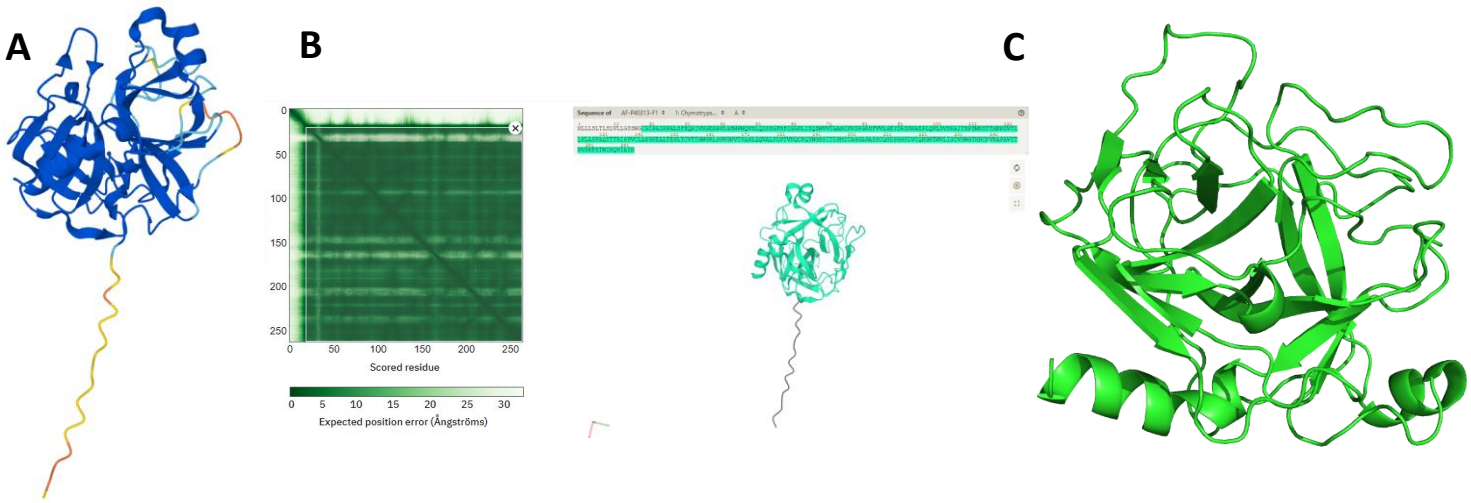


Figure 7. Key features of the ALPHAFOLD model generated by the pipeline A: ALPHAFOLD's pipeline estimated model of P40313 **B:** Confidence score visualised via ALPHAFOLD predicted aligned error and sequence view highlighted from amino acid 19 onwards **C:** ALPHAFOLD P40313 Desired Extracted Domain in green (without the low confidence score signal peptide from amino acids 1-18). Generated in *Pymol*.

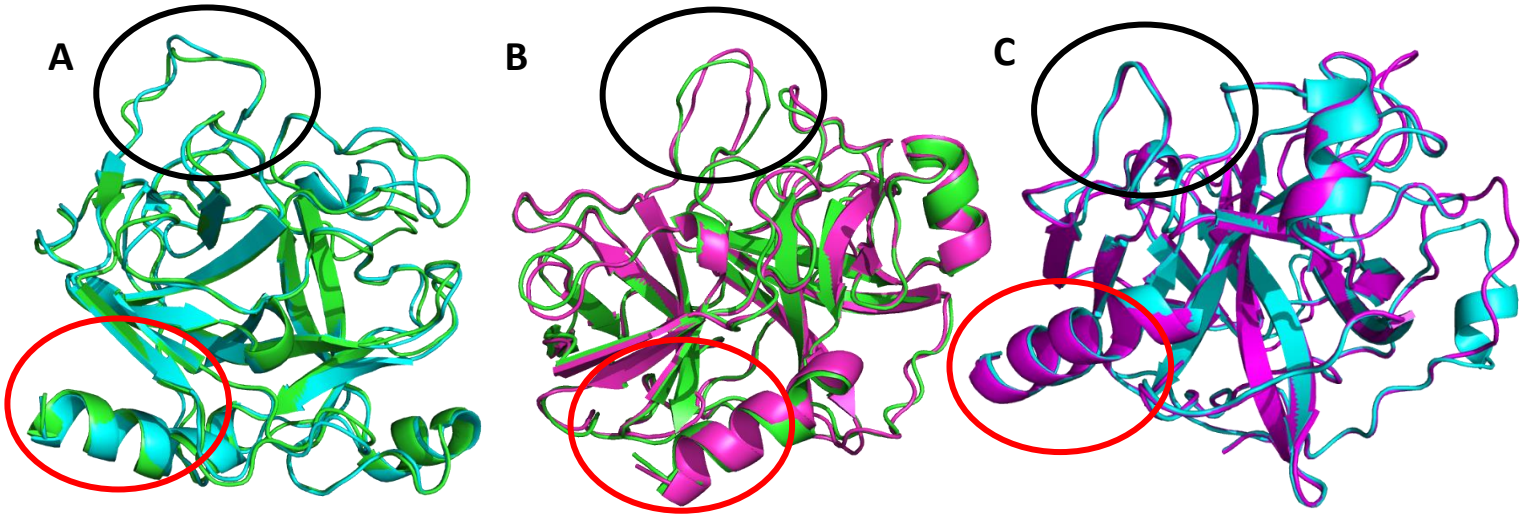


Figure 8. Superimposed Models, looped structure circled in black and alpha helices/B sheets circled in red. 1ACB_E alignment homology modelling. 1p2m.3.C automated homology modelling. A: ALPHAFOLD2 (green) superimposed with *1p2m.3.C* (blue) from SWISS-MODEL **B:** ALPHAFOLD (green) superimposed with *1ACB_E* (pink) from T-COFFEE. **C:** *1p2m.3.C* (blue) superimposed with *1ACB_E* (pink). Generated in *Pymol.*

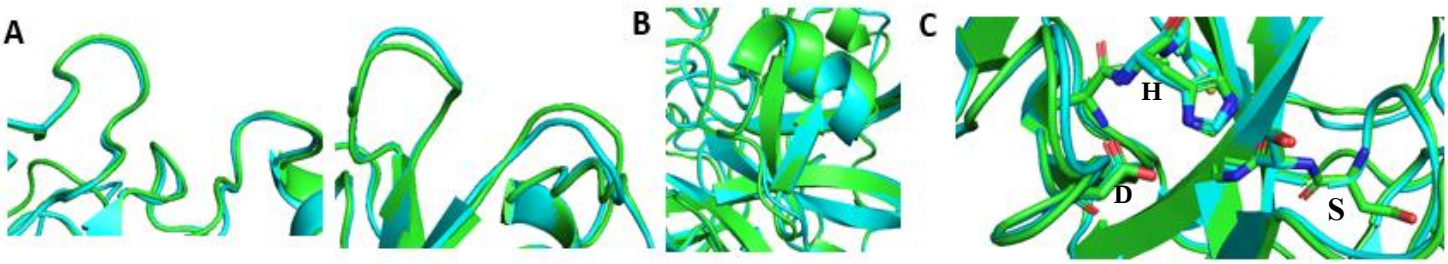


Figure 9. 1p2m.3.C (automated homology modelling - blue) superimposed with 1ACB_E (alignment homology modelling - green) snapshots: A: Loop structure snapshots **B:** Alpha helix snapshot **C:** The active site of the chymotrypsin-like protease (labelled - H: Histidine, D: Aspartate, S: Serine). Generated in *Pymol.*

b) Evaluate the model(s) from the parameters provided in the output of SWISS-MODEL. (max. 300 words)

High query coverage of templates meant most regions' coordinates could be mapped. RMSD can be misleading due to rigid body movement of domains or subdomains. The SWISS-MODEL metrics were obtained automatically by SWISS-PLOT in the homology modelling and manually for ALPHAFOLD via the structural assessment. The QMEANDisCo global score is a geometrical/global absolute quality estimator to measure consistency of the model with structural features predicted from the sequence. It is measured on a scale between 0 and 1 where a higher value is superior; the templates were very similar at 0.81 (**fig10**). However, the ALPHAFOLD2 was only 0.75 which is quite low but could be skewed by the extra signal peptide domain with no existing experimental structure at the protein size: the distance constraint (DisCo) score within assesses the agreement of pairwise distances in a model and an ensemble of constraints derived from experimentally determined protein structures that are homologous to the model (**fig11**) [Studer *et al.,* 2019]. The GMQE scores were very similar to the QMEANDisCo scores of the templates (0.812 and 0.804) as this primarily depends on coverage dependence which was the same for the sequence: the ALPHAFOLD has 100% coverage and thus did not have a score for this parameter.

The QMEAN score was the most polarising between the models. It is the primary assessor of model quality and has a -4.0 low quality threshold. A QMEANZ score of 0 indicates a very good geometrical property of a model to the experimental determined structure. The modelled *1ACB_E* template has by far the lowest QMEAN score at -0.04 (1p2m.3.c is -0.23), giving it the highest quality score; this is expected of alignment homology modelling which produces higher quality models than automated homology modelling in theory. Overall, the *1ACB_E* model is most accurate model although it still does not satisfy every statistic perfectly.

(max. 300 words)

Figure 10 The SWISS-MODEL outputs of the automated and alignment homology modelling estimations A: This shows the parameters of the SWISS-MODEL generated from *1ACB_E* by alignment homology modelling **B:** This shows the parameters of the SWISS-MODEL generated from *1p2m.3.C* by automated homology modelling.

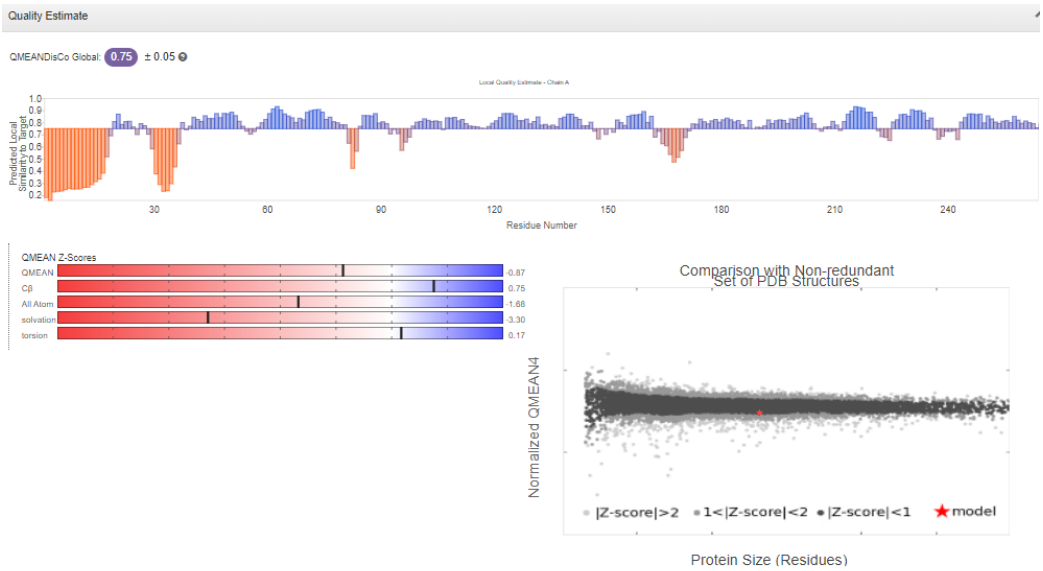


Figure 11: The SWISS-MODEL parameters of the ALPHAFOLD PDB. Obtained by a structure assessment of the model via SWISS-MODEL.

9) *Evaluate the model(s) and template structures with Molprobity.*

Display results from Molprobity. Discuss these results in light of your structural analyses, as well as your results on the template searches.

None of the Ramachandran general case plots (**fig12**) produced are perfect but they all positively evaluate the model quality and corresponding templates indirectly. The plot is used for comparative modelling via statistical distribution of backbone dihedral angle combinations: φ (x-axis) and ψ(y-axis). The top left is the B-strand, middle left the right-handed alpha-helix and right the L-protein enantiomers; however, glycine's plot has perfect symmetry due to loss of chirality via the hydrogen R group.

In theory, alignment homology modelling (*1ACB_E*) is generally favoured over automated homology modelling (*1p2m.3.C*) for model quality. This is generally supported by the plots as regardless of *1ACB_E's* slightly worse resolution (2.00Å rather than 1.75Å), its model has no amino acids in the disallowed region and most of its amino acids within the favourable regions – only 2 amino acids in the allowed region (**fig12.A**). *1p2m.3.C* has 4 amino acids in the allowed region and even a non-glycine amino acid, 224 Asn, in the disallowed region (**fig12.B**). The theoretical model produced by *ALPHAFOLD* is very low resolution due to its theoretical structure and it has two none-glycine residues in the white disallowed region which is a bad for structural reliability: these amino acids have impossible phi and psi angles within the structure (**fig12.C**).

Regardless of the minimal outliers, the data for bad bonds and bad angles is compromising. The goal for these data points is 0-0.1% but both templates sit in the red region. However, this is not due to the modelling process as the original experimentally determined templates like *1ACB_E* also have bonds and angles in the red regardless of no outliers (**fig13.B**). Overall, outliers are more important in assessing model quality so they can still be regarded as good models. Even the well-studied *KcsA* PDB (**fig13.A**) has far more outliers than our template models which shows our templates' quality in general. There is no 'perfect' structure and many plots still have many parameters labelled in red: red parameters are a warning and not impossible.

(max. 350 words)

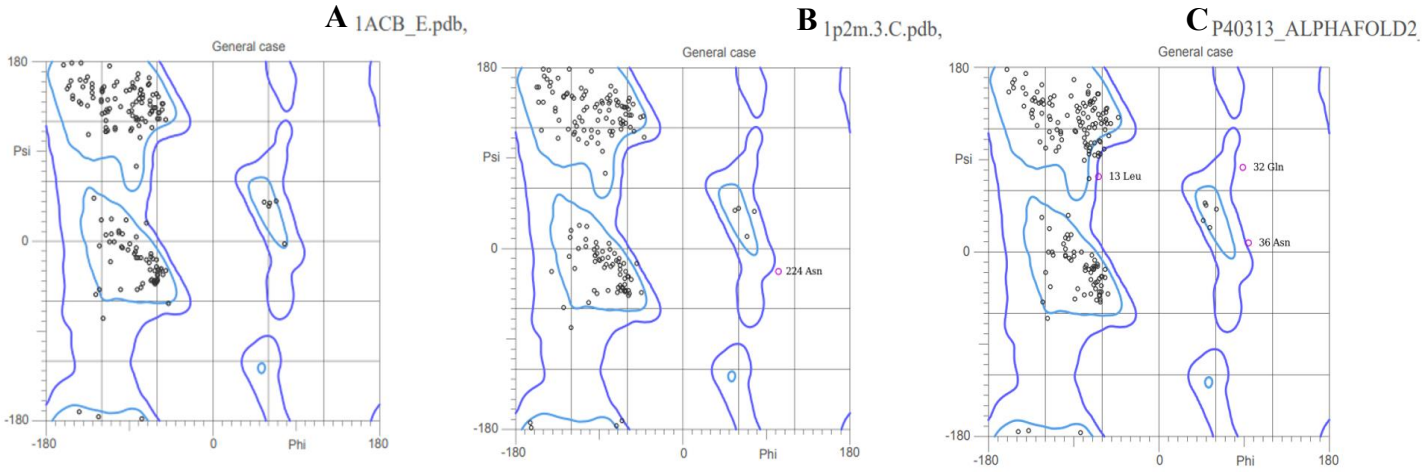
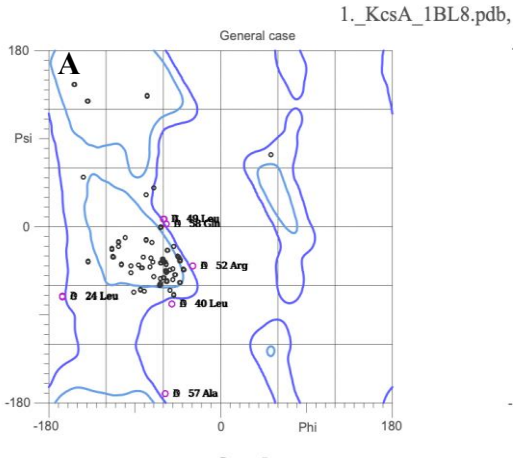


Figure 12. Ramachandran plots. The regions range from disallowed (white space) to allowed (enclosed by dark blue lines) and finally the most favourable regions (enclosed by light blue lines). A: Ramachandran plot of the alignment homology model from 1ACB_E produced by SWISS-PLOT using the T-COFFEE fasta_aln file **B:** Ramachandran plot of the automated homology model from *1p2m.3.C* **C:** The Ramachandran plot of the ALPHAFOLD2 model of P40313.



B

Protein Geometry	Poor rotamers	0	0.00%	Goal: <0.3%
	Favored rotamers	195	95.59%	Goal: >98%
	Ramachandran outliers	1	0.41%	Goal: <0.05%
	Ramachandran favored	230	94.26%	Goal: >98%
	Rama distribution Z-score	-1.23 ± 0.52		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	4	1.79%	Goal: 0
	Bad bonds:	1 / 1878	0.05%	Goal: 0%
	Bad angles:	27 / 2571	1.05%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 12	0.00%	Expected: ≤1 per chain, or ≤5%
	Cis nonProlines:	1 / 233	0.43%	Goal: <0.05%
Additional validations	Chiral volume outliers	0/297		

Figure 13. A: Ramachandran plot of the experimentally determined and crystallised KcsA Potassium Ion Channel for an impartial comparison. **B**: **The summary statistics** of the 1ACB_E template PDB.

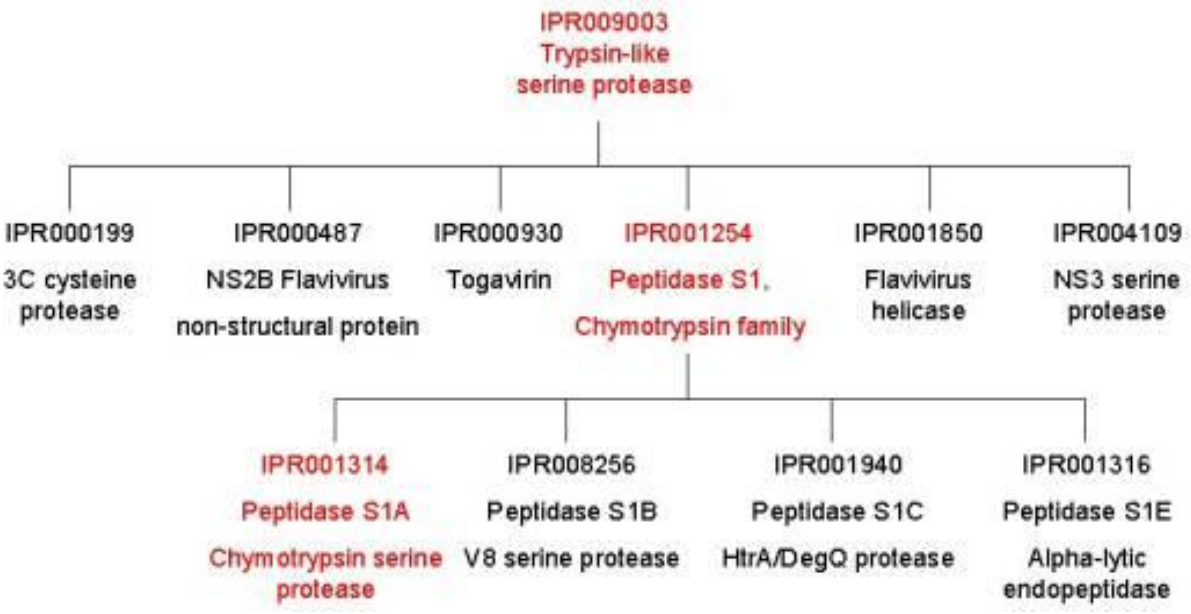


Figure 14. Table of chymotrypsin and trypsin proteins derived from INTERPRO This table shows family tree of trypsin-like serine proteases. The subfamilies corresponding to Peptidase S1 and Peptidase S1A are highlighted in red. [*PROTEINSWEBTEAM,* 2022]

10) Search for the functions of your assigned protein.

Using public repositories and servers, as well as any literature you have found, comment on possible functions for the gene/protein corresponding to the selected sequence.

The CTRL-1 gene in homo-sapiens has 8 transcripts and encodes a 264 amino acid serine-type endopeptidase with chymotrypsin and elastase-2-like activities: the target protein, P40313 [*Refseq*, 2016]. The serine-protease function is facilitated by a peptidase domain from amino acids 34-262 which is slightly longer than a trypsin domain (34-257), hence the classification of P40313 as a chymotrypsin [*Pfam*, 2022]. The target protein has a chymotrypsin-like geometric fold conserved evolutionarily which facilitates their unique c-terminal peptide-bond cleavage at aromatic residues [*INTERPRO*, 2022] [Ma *et al,* 2005]. This fold positions 3 essential active site residues: histidine (base), aspartate (nucleophile) and serine (nucleophile) of a catalytic triad (at positions 75, 121 and 214 respectively) charge relay system [Rawling and Barret, 1994]. The target chymotrypsin-like protease is specifically a serine-type endopeptidase in molecular function and hydrolyses internal alpha peptide bonds (**fig15,** highlighted).

The peptidase domain in the target protein is part of a subfamily S1A (a member of the MEROPS Peptidase family which can be extended at the N-terminus ((S1), **fig14**). This family includes the Chymotrypsin A from *Bos Taurus*, which was used as the template *1p2m.3.C* for modelling of the homo-sapiens target [*INTERPRO*, 2022]. The target protein is a zymogen and is predominantly expressed in the small intestine and pancreas for digestive function, regulated by its C-terminal pro-peptidase activation domain extension (amino acids 19-34); activation via cleavage of this pro-peptide domain is essential for the zymogen to function as it is released inactive in the pancreas to prevent self-digestion. The unified pro-peptide and peptidase domains in the target sequence are identified as a member of the TF330455 family via the curated ortholog database, linked to the chymotrypsin-like elastase family 2A which explains the elastase-2-like activities of the protein [*TreeFam*, 2022]. Furthermore, there is an additional c-terminal signal peptide domain from amino acids 1-18 in the target sequence for translocation that is absent in the target's BLAST homology templates from other species like *Bos Taurus*.

(max. 300 words)

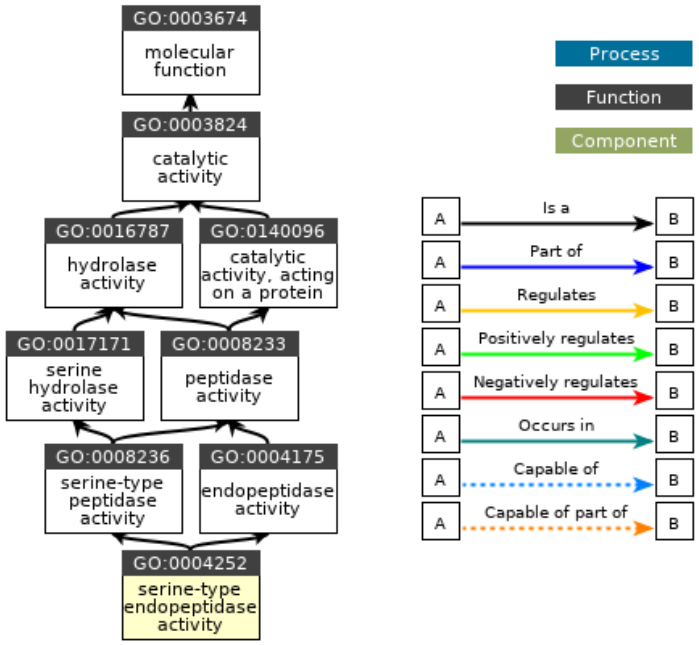


Figure 15: The serine-endopeptidase ancestry chart [*INTERPRO*, 2022]

THIS IS THE END OF THE IN-COURSE ASSIGNMENT

Make sure you have added essential plots, figures and analysis results in the essay.
You may add the alignments as an Appendix.

Appendix

References

1. Uniprot. (2022). UniProtKB - P40313 (CTRL_HUMAN)
2. Families of serine peptidases: Rawlings, N D, and A J Barrett. “Families of serine peptidases.” Methods in enzymology vol. 244 (1994): 19-61.
3. Interpro. (2022). Peptidase S1A, chymotrypsin family: http://www.ebi.ac.uk/interpro/entry/InterPro/IPR001314/
4. TreeFam. (2022). The phylogenetic tree of the TF330455 family
5. Ma W, Tang C, Lai L. Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant. Biophys J. 2005;89(2):1183-1193. doi:10.1529/biophysj.104.057158
6. Proteinswebteam. (2022). INTERPRO blog. https://proteinswebteam.github.io/interpro-blog/potm/2003_5/Page3.htm
7. RefSeq. (2016). ‘Summary’ in NCBI. https://www.ncbi.nlm.nih.gov/protein/NP_001898
8. Gabriel Studer, Christine Rempfer, Andrew M Waterhouse, Rafal Gumienny, Juergen Haas, Torsten Schwede, QMEANDisCo—distance constraints applied on model quality estimation, Bioinformatics, Volume 36, Issue 6, 15 March 2020, Pages 1765–1771

Alignments

1) The FASTA SEQUENCE OF THE TARGET

>sp|P40313|CTRL_HUMAN Chymotrypsin-like protease CTRL-1 OS=Homo sapiens OX=9606 GN=CTRL PE=1 SV=1
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFCGGSLISQSWVVT
AAHCNVSPGRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNNDVTLLKLASPAQYTTRISPVCLAS
SNEALTEGLTCVTTGWGRLSGVGNVTPAHLQQVALPLVTVNQCRQYWGSSITDSMICAGGAGASSCQGD
SGGPLVCQKGNTWVLIGIVSWGTKNCNVRAPAVYTRVSKFSTWINQVIAYN

2) The T-Coffee template 1 (1ACB_E) alignment FASTA

>query
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFCGGSLISQSWVVT
AAHCNVSP
GRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNNDVTLLKLASPAQYTTRISPVCLASSNEALTEGL
TCVTTGW
GRLSGVGNVTPAHLQQVALPLVTVNQCRQYWGSSITDSMICAGGAGASSCQGDSGGPLVCQKGNTWVLI
GIVSWGTKNCN

VRAPAVYTRVSKFSTWINQVIAYN
>1ACB_E

CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTS-
DVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVT
TGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGAW
TLVGIVSWGSSTCSTSTPGVYARVTALVNWVQQTLAAN

3) The T-Coffee template 2 (1DLK_B) alignment FASTA

>query
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFCGGSLISQSWVVT
AAHCNVSPGRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNNDVTLLKLASPAQYTTRISPVCLAS
SNEALTEGLTCVTTGWGRLSGVGNVTPAHLQQVALPLVTVNQCRQYWGSSITDSMICAGGAGASSCQGD
SGGPLVCQKGNTWVLIGIVSWGTKNCNVRAPAVYTRVSKFSTWINQVIAYN
>1DLK_B
----------IVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTT
S-
DVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVT
TGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGAW
TLVGIVSWGSSTCS
TSTPGVYARVTALVNWVQQTLAAN

3) The 1p2m.3.C alignment FASTA

>query
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFCGGSLISQSWVVT
AAHCNVSPGRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNNDVTLLKLASPAQYTTRISPVCLAS
SNEALTEGLTCVTTGWGRLSGVGNVTPAHLQQVALPLVTVNQCRQYWGSSITDSMICAGGAGASSCQGD
SGG
PLVCQKGNTWVLIGIVSWGTKNCNVRAPAVYTRVSKFSTWINQVIAYN
>template
----------CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVT
AAHCGVTT-
SDVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCV
TTGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGA
WTLVGIVSWGSSTCSTSTPGVYARVTALVNWVQQTLAAN

4) The 2cga.1. alignment FASTA

>query
MLLLSLTLSLVLLGSSWGCGIPAIKPALSFSQRIVNGENAVLGSWPWQVSLQDSSGFHFCGGSLISQSWVVT
AAHCNVSP
GRHFVVLGEYDRSSNAEPLQVLSVSRAITHPSWNSTTMNNDVTLLKLASPAQYTTRISPVCLASSNEALTEGL
TCVTTGW
GRLSGVGNVTPAHLQQVALPLVTVNQCRQYWGSSITDSMICAGGAGASSCQGDSGGPLVCQKGNTWVLI
GIVSWGTKNCN
VRAPAVYTRVSKFSTWINQVIAYN
>template

CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTS-
DVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVT
TGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGAW
TLVGIVSWGSSTCSTSTPGVYARVTALVNWVQQTLAAN