A benchmarked in-silico driven pipeline capable of complete trifecta protein analysis

Duwage Dahampriya Baladeva Deshan Perera

University of Sri Jayewardenepura <https://orcid.org/0000-0003-4654-0734>

Dinithi Peiris (✉ dinithi@sci.sjp.ac.lk)

<https://orcid.org/0000-0002-4788-6852>

Methodology article

Keywords: in-silico analysis, homology modelling, ab-initio modelling, virtual screening, molecular docking

Posted Date: February 7th, 2020

DOI: <https://doi.org/10.21203/rs.2.22948/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Current in-silico based proteomics requires the prediction, validation and functional analysis of a modelled protein. The main drawback in this protocol is the lack of a singular protocol that utilizes a proper set of benchmarked open science tools to accurately predict a protein's structure and function.

Results

The present study aims to utilize a series of openly available computational tools to formulate a complete pipeline capable of analyzing a novel protein's amino acid sequence, reveal its phylogenetic relationships and conserved residues. The resultant data would then be utilized in the formulation of the protein's tertiary structure followed by functional analysis through virtual screening. This novel protocol was benchmarked utilizing the Cystatin C protein obtained from *Danio rerio* and the AChE proteins of *Homo sapiens* and *Rattus norvegicus*.

Conclusion

The resultant pipeline utilizes freely available in-silico tools capable of complete protein analysis including a novel protein's functional analysis. The methodology has been proven and the tools have been benchmarked as the most suited for complete trifecta protein analysis.

Background

In the scientific endeavor of understanding the mechanisms of an organism's biology, the study of proteins, a key element in a number of cellular activities is vital. Proteins play both structural and functional roles within the organization of a cell [1]. The formation of a protein's characteristic and functionally related 3D structure begins from a simple but unique amino acid sequence [2].

The understanding of a novel amino acid sequence coding for an original protein begins with the identification of the sequence's relatedness to other proteins with a similar polypeptide chain [2]. This is followed by characterization of the protein and functional analysis through the study of its three-dimensional tertiary or even quaternary structure. For this three-level process of identification, characterization and function prediction, a plethora of bioinformatic tools are present [3]. But a singular protocol which encompasses all three processes utilizing the best suited tools based on efficiency and accuracy is lacking [4]. The present study aims to fill this gap by utilizing a series of opensource in-silico driven tools to achieve this goal in an economical manner and thus creating a tested and proven pathway for sound protein analysis.

With the completion of the human genome project and the genomes of entire organisms being decoded almost every day, the gap between proteins with known sequences and those with experimentally validated structures and function increase rapidly [5]. Therefore, the requirement of such a protocol, capable of converting genomic data into functional information has become more crucial [3]. For the determination of a protein's tertiary structure, a number of experimental procedures exist today; X-ray crystallography and Nuclear Magnetic Resonance are two of the most commonly used pathways today. But the primary shortcomings of these

methods are that they lack efficiency at the same time being highly costly [1]. This disables a major portion of the scientific community from conducting comprehensive studies in relation to a protein's structure and function primarily due to the inaccessibility to such specialized equipment and the cost.

The process of protein identification is the primary step of any protein study. The Basic Local Alignment Search Tool (BLAST) at the National Center for Biotechnology (NCBI) is a proven tool for protein identification. This process of sequence alignment provides the first connection between the DNA or amino acid sequence and the translated protein [6].

Following protein identification, the recognition of its relatedness to other closely related proteins via conserved sequence motifs and sequence similarity is crucial [7]. This step of the analysis helps reveal conserved functional domains. These conserved regions across species have been recognized through experimentation. They contribute to the macromolecules' function [8]. In-silico tools in regard to Phylogenetic analysis and Multiple Sequence Alignment (MSA) cater to this requirement.

On identification of a novel protein followed by characterization of its polypeptide chain, the next step is functional analysis [9]. In bioinformatics this process is twofold. The primary step requires the production of a structurally sound model of the protein's tertiary structure. This is followed by virtual screening of the protein for its functional effectiveness in the formation of stable protein-substrate complexes [5].

With the advent of modern medical applications such as personalized medicines and gene delivery, investigation into a protein's therapeutic potential has increased. This process typically requires a brute force endeavor of searching related literature and experimentation linking clinically significant pathways to the target macromolecule [10]. The in-silico driven approach of gene-gene interaction mapping is ideally catered to make this undertaking both efficient and reliable [11].

With the advancement in DNA sequencing and the increase in predicted amino acid sequences, a dire need for the comprehensive analysis of these novel proteins exists in the study of proteomics. The complete analysis as proven through experimentation, requires the identification of the amino acid sequence across a non-redundant database, followed by characterization and functional prediction through protein modelling techniques [12]. But in the present environment of bioinformatics, a singular, open-source and complete pipeline that caters to all of these functions is absent.

The present study attempts to fill this gap utilizing three instances. The first instance involves the amino acid sequence of *Danio rerio* which codes for the protease inhibiting Cystatin C protein and its interactions with human Cathepsin proteases. The second and third involve the Acetylcholinesterase (AChE) protein coding sequences of that of *Homo sapiens* and *Rattus norvegicus* and their interactions with the organophosphate Echothiophate.

Zebrafish commonly known as *Danio rerio* is a model organism of vertebrate development and subject to a number of scientific interests, especially in areas of modelling human biology and disease [13]. Cystatin C is a Type 2 Cystatin protein, present in all vertebrate organisms and known for being a competitive protease inhibitor of papain-like proteases such as Cathepsin B, H, L and S. They competitively block the cysteine active site of the Cathepsin proteins [14]. Due to their function as competitive inhibitors of proteases, they play a

crucial role in the regulation of certain diseases caused by cathepsin over expression such as atherosclerosis and metastasis of cancer cells [15].

Organophosphates such as Echothiophate are used to inhibit AChE activity. Their mechanism of AChE inhibition is well known and therefore makes them an ideal candidate in benchmarking the proposed pipeline. The study has focused on the AChE proteins of *Homo sapiens* (hAChE) and *Rattus norvegicus* (rAChE). AChE proteins have been predicted to possess an active site gorge with the presence of a tryptophan residue and Echothiophate is known to bind at this active site [16]. To successfully identify the evolutionarily conserved active sites these features should be looked out for [17].

The hAChE macro molecule has already been modelled and is readily available in the PDB data bank [18]. The primary reason for selection of this hAChE is to quantify the accuracy of the proposed pipeline in prediction of the actual structure. In contrast the rAChE molecule provided us with another instance to predict another novel protein and comparatively analyze it against its human counterpart.

The primary focus in the selection of the Cystatin C protein encoded by zebrafish as well as the AChE proteins was not only to benchmark the timeline with a multitude of queries, but to provide a wealth of information to the scientific community as well. The selection of an immune related protein provided the opportunity to benchmark all the features of the proposed pipeline. Here we report on novel in-silico tools to generate valid protein models, prove their functionality through molecular docking techniques and discover the immune potential against diseases in the human body through the use of gene-gene interaction mapping. The research study aims to develop an economical in-silico based approach to analyze the structure and functionality of novel proteins with a comprehensive revelation of their clinical significance.

Results

Identification of Conserved Domains

The initial analysis of the Cystatin C amino acid sequence obtained from the NCBI database (Accession No. AAZ29462.1) contained a Cystatin protein coding domain complete with a N-terminal glycine and a QXVXG sequence motif.

The analysis of the AChE of rAChE (AAH94521.1) and hAChE (AAA68151.1) amino acid sequences showed the presence of an Abhydrolase superfamily complete with a substrate binding pocket and a catalytic triad. AChE is a member of the Abhydrolase superfamily also referred to as Alpha/Beta hydrolase fold family of enzymes [19].

Phylogenetic data analysis

The resulting phylogenetic tree revealed that the query sequence did cluster itself among Type 2 Cystatin C proteins with a close resemblance to the Type 2 Cystatin protein produced by *Oncorhynchus keta* as depicted in Fig. 2.

In the instance of AChE, it could be observed that the proteins grouped themselves based on the phyla of Mammalia, Aves, Reptilia and Pisces. The AChE of humans and rats showed to share a common origin. The

overall phylogenetic tree was sound with confidence levels of over 80%.

Identification of conserved sequence motifs

The Cystatin C protein produced by *D. rerio* possessed the three evolutionary conserved sequence motifs, namely, the N-terminal Glycine, QVVAG motif and the PW motif [20]. *D. rerio* was recognized to contain a mutation in the PW motif which instead read as the LW motif as shown in Fig. 3.

In contrast, the AChE sequences for the mammalian clade had a high degree of conservation in relation to sequence motifs. The MSA also shows the high degree of similarity shared between AChE of *Rattus norvegicus* and *Homo sapiens*

Secondary structure prediction

The JPred server stated that the amino acid sequence under study coded for one short and one long alpha helix complete with five antiparallel beta sheets situated between the two helix structures. This was in compliance with a typical Cystatin C protein [21].

Both AChE amino acid chains had 14 alpha helices and 12 beta sheets. This data coincided with previous work conducted on AChE secondary structure prediction [22].

Protein structure prediction and refinement

The I-TASSER predicted structure was complete with five anti-parallel beta sheets lying across a short and long alpha helix complete with two disulfide bridges whereas the ab-initio predicted structure was devoid of both the short alpha helix and the disulfide bridges. Therefore, the homology model was selected while the later was rejected.

Quality assessment of the homology predicted structure through Pro-SA web service resulted in a Z-score of -4.55 which was nested within other proteins of similar size while the local model quality assessment revealed a series of N-terminal amino acids with unfavorable, elevated knowledge-based energies. Stereochemical analysis of the model through a Ramachandran plot revealed a mere 67.5% of the amino acid residues occupying the most favourable region. Due to the inadequacies in the structure which were revealed through these tests, it was decided to subject the homology predicted model to structure refinement using the ReFOLD server.

The ReFOLD server produced a high-resolution structure of 1.5 Å complete with the five antiparallel beta sheets, a short and long alpha helix and the two disulfide bridges characteristic to all Cystatin C proteins. The refined structure is depicted in Fig. 4.

The refined structure had a confidence score 6.31×10^{-8} which translated to an accuracy level of 99.99%. The ReFOLD server scored the refined structure a satisfactory global model quality score of 0.6671, stating that an overall improvement of 0.3% was performed on the final structure. Quality assessment of the refined structure using the ProSA-web service revealed an improved Z-score of -4.87 including a significant reduction in the knowledge-based energies of the N-terminal residues into favourable values. The final stereochemical analysis of the refined structure through the Ramachandran plot revealed that 76.1% of the amino acid residues occupied the most favourable regions while only 3.4% or four residues occupied the disallowed regions of the

plot. The results of the ProSA-web service and the Ramachandran plot of the refined structure are shown in Fig. 5. and Fig. 6.

The initially predicted AChE molecules showed certain discrepancies when subjected to the Ramachandran plot and ProSA web server analysis that were refined in the same manner using the ReFOLD server resulting in more stereochemically sound models.

Comparative analysis of predicted structure with existing structures

The prediction of the hAChE structure was conducted to analyze the degree of accuracy with which the predicted protein would match the existing model of the hAChE protein. The two structure did not perfectly match and there was some degree of variation. The RMSD showing the deviation of the two structures stood at 0.609.

Placement of evolutionarily conserved sequence motifs

The amino acid sequences encoding for Cystatin C proteins contains three main evolutionarily conserved sequence motifs namely the N-terminal Glycine, QVVAG motif and a PW motif. The positioning of these sequences on the proteins structure is directly related to its functional effectiveness [23]. The N-terminal glycine occupies the protein chain while the QVVAG motif and the PW motif inhabits the beta hairpin turns (Fig. 7). This structural placement proved that the predicted structure of the Cystatin C protein produced by *D. rerio* was functionally active [24, 25].

Surface analysis of tertiary structure

Analysis of the protein's surface for potential active sites revealed a total of 28 regions capable of forming viable protein-protein interactions. Twenty regions consisted of electrostatically charged patches resulting from a cumulation of six positively charged regions and 14 negatively charged regions. The remaining eight regions were hydrophobic in nature.

hAChE was revealed to possess 109 regions capable of protein-ligand interactions with 34 positive, 52 negative interactions and 21 hydrophobic interactions while the rAChE structure possessed 141 regions in total with 54 being positive, 62 negative and 23 hydrophobic interactions.

Functional analysis via virtual screening

The ability of the predicted Cystatin C structure to inhibit the activity of a cysteine protease enzyme was evaluated through rigid body protein-protein docking. The Cystatin C protein was first subjected to the inhibition of the cysteine protease Papain (PDB ID: 9PAP). A stable enzyme substrate complex was formed with a binding energy of -347.6 kJ/mol and the complete blockage of the ²⁵Cys and ¹⁵⁹His active sites of Papain [26].

Inhibition of Cathepsins revealed promising results. Cathepsin B's active sites of ²⁹Cys, ¹¹⁰His and ¹¹¹His were successfully inhibited with a binding energy of -542.3 kJ/mol [27]. Cathepsin H activity was inhibited with a binding energy of -601.8 kJ/mol with complete blockage of the ²⁵Cys and ¹⁵⁹His active sites [28]. Cathepsin L1 had the strongest protein-inhibitor complex with a binding energy of -793.4 kJ/mol. The Cathepsin L1 active sites of ²⁵Cys and ¹⁶³His showed complete blockage [29]. Finally, the activity of Cathepsin S was inhibited with

a binding energy of -435.7 kJ/mol and complete blockage of the active site ²⁵Cys [30]. The resultant protein-protein complexes produced by the virtual screening exercise are depicted in Fig. 9.

Virtual screening analysis for protein-ligand interactions

The protein ligand interactions were predicted using two different software, namely the Glide and AutoDock Vina. The predicted structures of rAChE and hAChE were subjected to virtual screening with Echothiophate. The purpose of using two different software enabled to identify if both software would in fact produce the same result.

The active site of AChE is situated in a centrally placed gorge [31]. The active site gorge contains a tryptophan residue involved in the formation of protein-ligand interactions [17]. The AutoDock Vina failed in the prediction of the active site gorge binding to the Echothiophate ligand in both instances with the rAChE and hAChE macro molecules. In contrast, the Glide Docking result showed that the Echothiophate organophosphate was nestled in the active site gorge with characteristic tryptophan residues being actively involved in the protein-ligand complex (Fig. 8).

Prediction of Cystatin C active binding site

Statistical analysis of the interacting residues of the Cystatin C protein responsible for the inhibition of cysteine protease activity revealed seven closely situated amino acids. These residues were identified as ²Phe, ³Leu, ⁹Phe, ¹²⁴Glu, ¹²⁵Asn, ¹²⁶Ser and ¹²⁷Cyx. The seven residues displayed their activity in the inhibition of the Cathepsin proteins. The predicted active binding site based on these findings is depicted in Fig. 10.

Human gene interaction mapping of Cathepsin pathways

Gene interaction mapping of Cathepsin B, Cathepsin L1 and Cathepsin S revealed a series of clinically significant pathways where Cathepsin over expression would lead to protein degradation and subsequent illness. The results of the gene interaction mapping are depicted in Fig. 11.

Cathepsin B overexpression resulted in the onset of osteoporosis, rheumatoid arthritis and certain forms of cancers due to the excessive degradation of the proteins Agrecan, Tenascin C, Fibronectin and Collagen type 1 [26, 32, 33]. Rigid body protein docking of Agrecan (PDB ID: 4M4D), Tenascin C (PDB ID: 2RB8), Fibronectin (PDB ID: 1E8B) and Collagen type (PDB ID: 3EJH) with Cathepsin B revealed successful enzyme substrate complexes with the respective binding energies of -263.2 kJ/mol – 500.1 kJ/mol, -551.8 kJ/mol and – 79.5 kJ/mol. All four of these binding energies except Fibronectin were greater than that of Cystatin C of *D. rerio* and Cathepsin B which stood at -549.2 kJ/mol.

Cathepsin L1 overexpression has proven to lead to the degradation of Fibronectin, resulting in the onset of melanomas [34]. Cathepsin L1 formed an enzyme substrate complex with Fibronectin with a binding energy of -508.7 kJ/mol which was greater than the binding energy of the Cystatin C protein under study and the Cathepsin L1 protein which stood at -793.4 kJ/mol.

Visualization of the gene interaction network of Cathepsin S revealed the degradation of Occludin which would result in the metastasis of cancer cells causing bone and breast cancers [35]. Protein docking of Cathepsin S with Occludin resulted in a low binding energy of -488.5 kJ/mol which was lower than that of the Cystatin C and Cathepsin S complex which stood at -454.4 kJ/mol.

Discussion

The primary objective of the present study is to design a novel procedure which attempts to identify, characterize and predict the structure of the protein under study, prove its functional capability and predict its therapeutic potential for the treatment of diseases. Even though this is the standard procedure for protein analysis, a complete singular article encompassing the steps in order with benchmarked tools was lacking.

Proteins belonging to the same family usually have a series of conserved sequence motifs that act as a fingerprint enabling the identification of novel sequences. The Cystatin C coding amino acid sequence of *Danio rerio* obtained from the NCBI database should have a Cystatin coding domain. The Cystatin Superfamily coding sequence motifs consists of a N-terminal glycine and a QXVXG sequence motif [36]. The Conserved Domain (CD) search revealed the presence of a Cystatin domain with the two evolutionary conserved sequence motifs. In the event a DNA sequence coding for a novel protein is present, it is recommended to commence the study using NCBI's BLASTx search [37].

Conserved segments in DNA or protein sequences play a role in the functioning of the macromolecules encoded by them [38]. Likewise, proteins of similar function belong to the same family. Therefore, the positioning of a predicted protein among others of shared functions and the presence of these evolutionarily conserved sequences would enable the validation of a predicted protein [39].

The phylogenetic construction of the Cystatin superfamily tree conducted by the MEGA 6 software revealed a Cystatin C dendrogram which coincides with the existing validated Cystatin protein family. The generated phylogenetic tree depicted in Fig. 2 had a common ancestry between the Type 1 Cystatins, Cystatin A and B, with a branching off to a shared ancestry between Type 2 and 3 Cystatins. A complete separate branch is shown by Type 4 Cystatins or Fetuins. This phylogenetic pattern which was successfully reproduced in the study is characteristic to the Cystatin superfamily [14]. The protein under study was found to be positioned among the Type 2 Cystatin C proteins. This confirms the accurate prediction of the phylogenetic tree.

Execution of a Multiple Sequence Alignment (MSA) post phylogenetic analysis enables the accurate recognition of evolutionarily conserved sequence motifs among closely related proteins that are truly involved in its characteristic functions. It also enables the identification of sequence motifs that are most critical in the governance of the protein's function [40]. Cystatin C proteins are supposed to have 3 such evolutionarily preserved sequence motifs, namely, an N-terminal Glycine, a QVVAG motif and a PW motif [14]. The MSA revealed the presence of all three sequence motifs being conserved in the polypeptide chain of the protein under study.

The preceding step after protein characterization is the process of functional analysis which requires the prediction of the protein's native 3D tertiary structure. In an iterative fashion this involves the identification of the protein's secondary structure first [2]. The secondary structure of a typical Cystatin C protein must have two

alpha helices, one short and one long alpha helix and between them, five antiparallel beta sheets [41]. All of these features were shown to be present by the JPred server in the amino acid sequence under study.

The prediction of the protein's tertiary structure through computational techniques has two-fold advantages. Conventional methods of NMR and X-Ray Crystallography are expensive and not readily available to the masses. Also, the structure of short half-life proteins cannot be predicted using these methods [1]. The current pipeline successfully utilizes the two main methods of protein modelling available, namely, homology and ab-initio modelling [42].

The I-TASSER server was utilized for the Homology prediction of the protein structure. The I-TASSER server undertakes a hierarchical approach to the prediction of protein structure [43]. Its reliability has been proven by the Critical Assessment of methods of protein Structure Prediction (CASP) experiments [44]. The predicted structure contained all the structural components that must be present in a Cystatin C protein such as both a short and long alpha helix, five anti parallel beta sheet and two disulfide bonds [21].

Ab-initio or de-novo protein modelling was conducted utilizing the QUARK server. In the absence of known templates this technique can be utilized to predict the protein structure from the amino acid sequence alone [42]. The QUARK server was selected due to its reliability proven through the CASP experiments [45]. But the ab-initio predicted model failed the validity test for both of its alpha helices were long and had no disulfide bonds.

The structure of a protein depends on its structure. A valid tertiary structure should be at a minimum free energy and be stereochemically stable [46]. These two tests carried out by the Pro-SA web test and the Ramachandran plots revealed the structure had a few discrepancies. In such an event, the refinement of the protein has proven to be an efficient solution. The ReFOLD server validated through the CASP12 experiments was capable of removing these discrepancies and showing significant improvements on retesting of the refined model [47].

It is known that electrostatically charged and hydrophobic residues on a protein's surface plays a key role in long distance protein-protein interactions and are vital for drug design. Therefore, these regions were mapped before the conduction of the virtual screening process [48]. The Cystatin C under study had 35 such regions.

Cystatin C proteins are dubbed to be the most active protease inhibitors from the entire Cystatin family. It is capable of competitively blocking the active site of Papain, Cathepsin B, H, L and S [14]. Protein-protein docking proved this statement to be accurate for in each instance the cysteine active site of the Cathepsin protein was blocked by the Cystatin C protein. Based on Gibbs free energy laws proteins forming valid complexes should have minimum negative values [1]. The inhibition of the Cathepsins all occurred under favourable negative binding energies further cementing the results of the pipelines chosen docking software.

Residues that continuously take part in the formation of the protein-protein complexes are believed to be related to the protein's function [4]. Such residues for the protein active site [49]. Utilizing this principal, the active site of the Cystatin C protein was identified by recognition of the amino acid sequences that were always involved in the inhibition of the Cathepsin proteins (Fig. 10).

Genetic interactions mapping reveals functional pathways involved in molecular mechanism [50]. Some of these mechanisms can become clinically significant if compromised and lead to disease conditions such as tumorigenesis. Therefore, the study of gene-gene interaction maps of model organisms and humans have been

incorporated in drug discovery and design [51]. Gene interaction mapping of Cathepsins B, L1 and S revealed a number of clinically compromised pathways caused by Cathepsin overexpression (Fig. 11). Considering the fact that the Cystatin C protein is a competitive inhibitor, cross analysis of the binding affinity of Cathepsins with that of their natural substrates and the Cystatin C protein should reveal the potential of the protein understudy to act as a potential drug [52]. The viable candidates were diseases caused by overexpression of Cathepsin B and L1.

The benchmarking of the pathway was led with the prediction of the structure and function of the Cystatin C of *D. rerio*. To further test the reliability and accuracy of the proposed pathway and the software used, the AChE proteins of both humans and rats was selected. This also enabled the extension of the pathway from being protein-protein to protein-ligand analysis as well.

The use of the hAChE macro molecule was to test the reliability of the I-Tassser software against already identified proteins structures. The hAChE protein's tertiary structure already exists in the PDB databank. Even though the server predicted a structure closely resembling the true native model after refinement there were minor deviations which resulted in the RMSD value being 0.609. Apart from this, both the models were successfully analyzed for their functionality.

The AChE protein was utilized to comparatively analyze the use of AutoDock Vina and Glide docking. The resultant data showed that in contrast to the Glide docking result that the AutoDock Vina failed to accurately predict the ligand binding site. Instead of the ligand forming a complex in the active site gorge it was bound to the surface of the protein. Thus, ensuring that the Glide docking software was more reliable in the prediction of protein-ligand interactions.

Conclusion

In conclusion, the proposed computational pipeline utilizing solely free, opensource software is able to conduct a complete analysis of a novel amino acid sequence revealing its identity, functional and therapeutic potential. The pipeline was able to characterize the protein based on its evolutionary relatedness, overcome protein modelling errors through refinement and reveal the functionality of the protein complete with its active site and therapeutic potential. Therefore, we believe that this singular, economical pipeline will be capable of assisting in bridging the gap between known protein sequences and those that are experimentally validated.

Methods

The proposed methodology's principal focus is the creation of a sound tertiary structure from an amino acid sequence (Fig. 1). The procedure was initially benchmarked utilizing the Cystatin C proteins obtained from *Danio rerio* followed by consolidation of the protocol through Human AChE (hAChE) and AChE of *Rattus norvegicus* (rAChE).

The use of these three proteins enabled the testing of the protocol to counter all possible situations that may arise in the structure and functional analysis of a protein. The Cystatin C of *D. rerio* and rAChE proteins enabled the protocol to conduct the trifecta analysis of novel structures and test their functions in both protein-protein and protein-ligand interactions. The hAChE instance enabled the methodology to be comparatively tested

against a protein of known structure. The true structure of hAChE is readily available in the PDB data bank (PDBID: 4PQE).

Identification of Conserved Domains

The amino acid sequence coding for the putative Cystatin C protein by *D. rerio* (Accession No. AAZ29462.1), rAChE (AAH94521.1) and hAChE (AAA68151.1) was identified using the NCBI database. The conserved functional domains present in the translated amino acid sequence were identified using the Conserved Domain Search Service (CD Search) in NCBI. The predicted protein was annotated and a graphical summary was obtained [53].

Phylogenetic data analysis

The evolutionarily related protein families to the predicted protein of *Danio rerio* was identified using the HUGO Gene Nomenclature Committee Search tool. The related proteins were searched for in the NCBI database and the amino acid sequence in the FASTA format was extracted. It was ensured that at least one amino acid sequence was obtained for each of the four vertebrate classes; Pisces, Amphibia, Reptilia, Aves and Mammalia. These amino acid sequences were subsequently aligned by ClustalW using the software MEGA6 and a phylogenetic tree was obtained using the neighbor-joining statistical method.

Identification of conserved sequence motifs

The translated protein sequences were subjected to MSA by ClustalW using the program Unipro UGENE [54]. with the MSA was conducted along with ten of its most closely related amino acid sequences in order to identify sequence motifs that have been evolutionarily conserved.

Amino acid sequence and secondary structure characterization

The secondary structure being expressed by the amino acid sequence was predicted using the JPred secondary structure prediction server [55].

Tertiary structure prediction

The 3D structure of the protein was predicted using two main systems. Homology prediction of the protein was conducted using the I-TASSER online software [5]. Ab-initio model prediction was conducted by the QUARK server [56]. Ab-initio modelling was not utilized for the prediction of the AChE molecules due to the over 200 amino acid limitation of the server. The predicted 3D structures were visualized using the PyMOL visualization software [57].

Cystatin C proteins are required to possess a series of evolutionary conserved structural features. The characteristic features include a short and long alpha helix lying across a five-stranded anti parallel beta sheet with two disulfide bridges. The predicted structure possessing all of these features was selected as the most accurate model.

The tertiary structure was AChE, consisting of a centrally placed mixture of beta sheets that are surrounded by 15 alpha helices [17].

Quality assessment of predicted structure

The successfully predicted structure was evaluated for its overall quality and stability in order to recognize any errors that may be present. The overall model quality and the local model quality was evaluated using the ProSA-web service [58]. The stability and stereochemistry of the structure was assessed by generating a Ramachandran plot using the PROCHECK software [59].

Tertiary structure refinement

The overall tertiary model quality was improved through the ReFOLD online software. The software was utilized to assess the global model quality, the overall structural improvement and accuracy [47]. The refined structure was then subjected to validation and quality assessment once more to ensure it met all the requirements of the protein.

Surface analysis of the tertiary structure

The surface of the protein structure was analyzed for potential active binding sites by examining the distribution of electrostatic charges and hydrophobic amino acids using the Protein Surface Analyzer tool from the Maestro BioLuminate 2.8 software [60].

Functional analysis through molecular docking

In the functional analysis of a protein it is essential to determine the type of interactions the target macro molecule will be taking part in. The following protocol attempts to tackle the most common forms of protein interactions namely protein-ligand and protein-protein interactions.

For the purpose of protein-protein interactions the Cystatin C protein along with its numerous Cathepsin substrates were considered. Rigid body protein-protein docking was conducted to analyze the functional effectiveness of the Cystatin C protein produced by *D. rerio* as a cysteine protease inhibitor. Hex 8.0.0 Cuda was used and Fast Fourier Transform (FFT) correlation techniques were involved [61]. The required protein models were obtained from the Protein Data Bank (RCSB). For successful inhibition on the cysteine protease the Cystatin C protein had to block the cysteine active site of the target protein with a favourable negative binding energy. The protein was first subjected to molecular docking with Papain (PDB ID: 9PAP), known to be inhibited by all Cystatins. This was followed suit by Papain like proteases namely Cathepsin B (PDB ID: 2IPP), Cathepsin H (PDB ID: 8PCH), Cathepsin L1 (PDB ID: 2Y2J) and Cathepsin S (PDB ID: 2FRQ). The proteases have been proven to be actively inhibited by Cystatin C [62]. The interacting amino acids of the Cystatin C protein in each protein-protein interaction was analyzed through Maestro BioLuminate 2.8 software.

In contrast to protein-protein interaction studies, it was observed that there are numerous programs available for the purpose of studying protein-ligand interactions. These vary from commercial to free, open source licensed software. In the study, two commonly used software from each category were selected. AutoDock Vina was selected to represent the open source software whereas Glide docking was utilized to represent the commercial workspace, but it should be noted that Glide docking is available for free on trial for students and researchers. The protein-ligand substrate consisted of AChE and the organophosphate Echothiophate.

Analysis of the therapeutic potential

Clinically significant pathways involving Cathepsin B, L1 and S over expression in the human body were identified through gene-gene interaction mapping. Cathepsin H was excluded due to the absence of its 3D

tertiary structure of human origin. The GeneMANIA database coupled with the Cystoscape software was used to identify these protein interactions. The interactions were isolated to physical interactions, co-expression and co-localization [63]. The selected natural substrates being subjected to degradation by Cathepsin over expression resulting in a clinical response were subjected to molecular docking with the respective Cathepsin and the binding energies were compared with that of Cystatin C produced by *D. rerio*. If the Cystatin C of *D. rerio* had a more feasible binding energy, it was estimated to have a clinical potential as a drug to solve the Cathepsin over expression.

Abbreviations

BLAST: Basic Local Alignment Search Tool; NCBI: National Center for Biotechnology; MSA: Multiple Sequence Alignment; AChE: Acetylcholinesterase; hAChE: AChE proteins of *Homo sapiens*; rAChE: *Rattus norvegicus*; CD search: Conserved Domain search; PDB: Protein Data Bank

Declarations

Competing interests

The authors declare no competing interests.

Author contributions

L.D.C.P is the supervising professor of the research project and was involved in the design and execution of the research work.

D.D.B.D.P conducted the research work and was involved with the design and execution of the research methodology.

Data availability

All the data and software utilized in the design and benchmarking of the pipeline are open source and freely available for download. The relevant software and accession numbers are mentioned in the manuscript text.

Additional Information

Supplementary information is provided

Acknowledgments

Not applicable

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Funding

Not applicable

References

1. Zarbafian S, Moghadasi M, Roshandelpoor A, Nan F, Li K, Vakli P, et al. Protein docking refinement by convex underestimation in the low-dimensional subspace of encounter complexes. *Sci Rep.* 2018;8:1–12.
2. Godbey WT. Proteins. In: *An Introduction to Biotechnology*. Elsevier; 2014. p. 9–33. doi:10.1016/B978-1-907568-28-2.00002-2.
3. Skarzyńska A, Pawełkowicz M, Krzywkowski T, Świerkula K, Pląder W, Przybecki Z. Bioinformatics pipeline for functional identification and characterization of proteins. 2015; October:96621M. doi:10.1117/12.2205559.
4. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep.* 2017;7:1–15. doi:10.1038/s41598-017-09654-8.
5. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* 2014;12:7–8. doi:10.1038/nmeth.3213.
6. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* 2013;41 Web Server issue:29–33.
7. Margulies EH. Identification and Characterization of Multi-Species Conserved Sequences. *Genome Res.* 2003;13:2507–18. doi:10.1101/gr.1602203.
8. Wong A, Gehring C, Irving HR. Conserved Functional Motifs and Homology Modeling to Predict Hidden Moonlighting Functional Sites. *Front Bioeng Biotechnol.* 2015;3 June:1–8. doi:10.3389/fbioe.2015.00082.
9. Makrodimitris S, van Ham RCHJ, Reinders MJT. Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics.* 2018; August:1–9. doi:10.1093/bioinformatics/bty751.
10. Gurevich E V, Gurevich V V. Therapeutic Potential of Small Molecules and Engineered Proteins. In: Gurevich V V, editor. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 1–12. doi:10.1007/978-3-642-41199-1_1.
11. Bebek G. Identifying Gene Interaction Networks. 2012. p. 483–94. doi:10.1007/978-1-61779-555-8_26.
12. Pruess M, Apweiler R. Bioinformatics Resources for In Silico Proteome Analysis. *J Biomed Biotechnol.* 2003;2003:231–6.
13. Clark KJ, Balciunas D, Pogoda H-M, Ding Y, Westcot SE, Bedell VM, et al. In vivo protein trapping produces a functional expression codex of the vertebrate proteome. *Nat Methods.* 2011;8:506–12. doi:10.1038/nmeth.1606.
14. Ochieng J, Chaudhuri G. Cystatin Superfamily. *J Health Care Poor Underserved.* 2010;21:51–70. doi:10.1353/hpu.0.0257.
15. Magister Š, Kos J. Cystatins in immune system. *J Cancer.* 2013;4:45–56.
16. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res.* 2002;30:412–5. doi:10.1093/nar/30.1.412.

17. Dvir H, Silman I, Harel M, Rosenberry TL, Sussman JL. Acetylcholinesterase: From 3D structure to function. *Chem Biol Interact.* 2010;187:10–22. doi:10.1016/j.cbi.2010.01.042.
18. Dym O, Unger T, Toker L, Silman I, Sussman JL, (ISPC) ISPC. Crystal Structure of Human Acetylcholinesterase. Protein Data Bank. 2015. doi:10.2210/PDB4PQE/PDB.
19. Holmquist M. Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr Protein Pept Sci.* 2000;1:209–35. <http://www.ncbi.nlm.nih.gov/pubmed/12369917>. Accessed 8 Jun 2019.
20. Paraoan L, Hiscott P, Gosden C, Grierson I. Cystatin C in macular and neuronal degenerations: Implications for mechanism(s) of age-related macular degeneration. *Vision Res.* 2010;50:737–42. doi:10.1016/j.visres.2009.10.022.
21. Kolodziejczyk R, Michalska K, Hernandez-Santoyo A, Wahlbom M, Grubb A, Jaskolski M. Crystal structure of human cystatin C stabilized against amyloid formation. *FEBS J.* 2010;277:1726–37.
22. Dvir H, Silman I, Harel M, Rosenberry TL, Sussman JL. Acetylcholinesterase: From 3D structure to function. *Chem Biol Interact.* 2010;187:10–22. doi:10.1016/j.cbi.2010.01.042.
23. Premachandra HKA, Wan Q, Elvitigala DAS, De Zoysa M, Choi CY, Whang I, et al. Genomic characterization and expression profiles upon bacterial infection of a novel cystatin B homologue from disk abalone (*Halotis discus discus*). *Dev Comp Immunol.* 2012;38:495–504. doi:10.1016/j.dci.2012.06.010.
24. Björk I, Brieditis I, Raub-Segall E, Pol E, Håkansson K, Abrahamson M. The importance of the second hairpin loop of cystatin C for proteinase binding. Characterization of the interaction of Trp-106 variants of the inhibitor with cysteine proteinases. *Biochemistry.* 1996;35:10720–6.
25. Lewandowska A, Ołdziej S, Liwo A, Scheraga HA. Beta-Hairpin-Forming Peptides; Models of Early Stages of Protein Folding. *Biophys Chem.* 2010;151:1–9. doi:10.1016/j.bpc.2010.05.001.
26. Fonović M, Turk B. Cysteine cathepsins and extracellular matrix degradation. *Biochim Biophys Acta - Gen Subj.* 2014;1840:2560–70.
27. Musil D, Zucic D, Turk D, Engh RA, Mayr I, Huber R, et al. The refined 2.15 Å X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. *EMBO J.* 1991;10:2321–30. <http://www.ncbi.nlm.nih.gov/pubmed/1868826>5Cn<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC452927>.
28. Gunčar G, Podobnik M, Pungerčar J, Štrukelj B, Turk V, Turk D. Crystal structure of porcine cathepsin H determined at 2.1 Å resolution: Location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function. *Structure.* 1998;6:51–61.
29. Gunčar G, Pungerčič G, Klemenčič I, Turk V, Turk D. Crystal structure of MHC class II-associated p41 li fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. *EMBO J.* 1999;18:793–803.
30. Mcgrath ME, Palmer JT, Brömme D, Somoza JR. Crystal structure of human cathepsin S. *Protein Sci.* 1998;7:1294–302. doi:10.1002/pro.5560070604.
31. Axelsen PH, Harel M, Silman I, Sussman JL. Structure and dynamics of the active site gorge of acetylcholinesterase: synergistic use of molecular dynamics simulation and X-ray crystallography. *Protein Sci.* 1994;3:188–97. doi:10.1002/pro.5560030204.
32. Mai J, Sameni M, Mikkelsen T, Sloane BF. Degradation of extracellular matrix protein tenascin-C by cathepsin B: An interaction involved in the progression of gliomas. *Biol Chem.* 2002;383:1407–13.

33. Li W, Liu Z, Zhao C, Zhai L. Binding of MMP-9-degraded fibronectin to α_6 integrin promotes invasion via the FAK-Src-related Erk1/2 and PI3K/Akt/Smad-1/5/8 pathways in breast cancer. *Oncol Rep*. 2015;34:1345–52. doi:10.3892/or.2015.4103.
34. Chakraborti S, Chakraborti T, Dhalla NS. *Proteases in Human Diseases*. Singapore: Springer Singapore; 2017. doi:10.1007/978-981-10-3162-5.
35. Martin TA, Jordan N, Davies EL, Jiang WG. Metastasis to Bone in Human Cancer Is Associated with Loss of Occludin Expression. *Anticancer Res*. 2016;36:1287–93. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=26977027>.
36. Dutt S, Singh VK, Marla SS, Kumar A. In silico Analysis of Sequential, Structural and Functional Diversity of Wheat Cystatins and Its Implication in Plant Defense. *Genomics, Proteomics Bioinforma*. 2010;8:42–56.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.
38. Stojanovic N, Florea L, Riemer C, Gumucio D, Slightom J, Goodman M, et al. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res*. 1999;27:3899–910. doi:10.1093/nar/27.19.3899.
39. Kordis D, Turk V. Phylogenomic analysis of the cystatin superfamily in eukaryotes and prokaryotes. *BMC Evol Biol*. 2009;9:266. doi:10.1186/1471-2148-9-266.
40. Jankun-Kelly TJ, Lindeman AD, Bridges SM. Exploratory visual analysis of conserved domains on multiple sequence alignments. *BMC Bioinformatics*. 2009;10 SUPPL. 11:3–11.
41. Abrahamson M, Alvarez-Fernandez M, Nathanson C-M, Withana NP, Ma X, McGuire HM, et al. Cystatins. *Biochem Soc Symp*. 2003;99:179–99. doi:10.1042/bss0700179.
42. Zhang Z. An Overview of Protein Structure Prediction: From Homology to Ab Initio Final Project For Bioc218, Computational Molecular Biology. <https://pdfs.semanticscholar.org/522a/f9cf5d1c3e4c1506d449286de6d3ebbd07ef.pdf>. Accessed 17 Dec 2018.
43. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 2010;5:725–38. doi:10.1038/nprot.2010.5.
44. Zhang Y, Arakaki AK, Skolnick J. TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins Struct Funct Genet*. 2005;61 SUPPL. 7:91–8.
45. Zhang W, Yang J, He B, Walker SE, Zhang H, Govindarajoo B, et al. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins Struct Funct Bioinforma*. 2016;84 Suppl 1:76–86. doi:10.1002/prot.24930.
46. Miklos AC, Li C, Pielak GJ. Using NMR-Detected Backbone Amide ^1H Exchange to Assess Macromolecular Crowding Effects on Globular-Protein Stability. *Methods Enzymol*. 2009;466:1–18. doi:10.1016/S0076-6879(09)66001-8.
47. Shuid AN, Kempster R, McGuffin LJ. ReFOLD: A server for the refinement of 3D protein models guided by accurate quality estimates. *Nucleic Acids Res*. 2017;45:W422–8.
48. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews*. 2008;108:1225–44. doi:10.1021/cr040409x.

49. Chatterjee A, Roy UK, Halder D. Protein Active Site Structure Prediction Strategy and Algorithm. *Int J Curr Eng Technol*. 2011; June. doi:10.14741/ijcet/22774106/7.3.2017.53.
50. Jaimovich A, Rinott R, Schuldiner M, Margalit H, Friedman N. Modularity and directionality in genetic interaction maps. *Bioinformatics*. 2010;26:228–36.
51. Saxena N, Saxena V. Gene-Gene Interaction Mapping Of Human Cytomegalic Virus through System Biology Approach. *Biol Syst Open Access*. 2015;04:2–7. doi:10.4172/2329-6577.1000141.
52. Lionta E, Spyrou G, Vassilatis D, Cournia Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr Top Med Chem*. 2014;14:1923–38. doi:10.2174/1568026614666140929124445.
53. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res*. 2015;43:D222–6.
54. Okonechnikov K, Golosova O, Fursov M, Varlamov A, Vaskin Y, Efremov I, et al. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*. 2012;28:1166–7. doi:10.1093/bioinformatics/bts091.
55. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: A protein secondary structure prediction server. *Nucleic Acids Res*. 2015;43:W389–94. doi:10.1093/nar/gkv332.
56. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct Funct Bioinforma*. 2012;80:1715–35.
57. DeLano WL. PyMOL: An Open-Source Molecular Graphics Tool. *CCP4 Newsl Protein Crystallogr*. 2002;40, 82–92. <http://www.ccp4.ac.uk/newsletters/newsletter36.pdf>.
58. Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*. 2007;35 SUPPL.2:407–10.
59. Laskowski RA, Macarthur MW, Thornton JM. Chapter 21 . 4 . PROCHECK: validation of protein-structure coordinates. 2012;:684–7.
60. Zhu K, Day T, Warshaviak D, Murrett C, Friesner R, Pearlman D. Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins Struct Funct Bioinforma*. 2014;82:1646–55. doi:10.1002/prot.24551.
61. Ritchie DW. Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins Struct Funct Genet*. 2003;52:98–106.
62. Kopitar-Jerala N. The role of cystatins in cells of the immune system. *FEBS Lett*. 2006;580:6295–301.
63. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38 SUPPL. 2:214–20.

Figures

Figure 1

A graphical representation of the proposed methodology summarizing the sequence of steps that should be followed. This the main pathway that should be followed and it is made flexible to in-silico tools so that the user may customize the tools used for each step.

Figure 2

Refined Phylogenetic tree. Type 1 Cystatins in Green, Type 2 Cystatins in Blue, Type 3 Cystatins in Red and Type 4 Cystatins in Orange. The *Danio rerio* amino sequence highlighted yellow.

Figure 3

Alignment of the *Danio rerio* amino acid sequence with the closely related sequences. The conserved G, QVVA motif and PW motif are highlighted in yellow. Highly conserved, strong homologies are shaded in dark blue and weak similarities shaded in light blue.

Figure 4

Refined tertiary structure model produced by ReFOLD server. The N-terminus (protein start) is coloured in orange, the C-terminus (protein end) is coloured in Blue, beta sheets are coloured in Yellow, alpha helixes are coloured in Red and disulfide bonds are coloured in Grey.

Overall model quality

Z-Score: **-4.87**

Local model quality

Figure 5

ProSA-web service quality assessment plots for Refined ReFOLD model. The overall model quality diagram on the left gives a comparative analysis of proteins of the same sequence length to that of the query protein's structure. The query protein is demarcated as a black dot and if it is in the vicinity of other proteins of similar size the structure is sound. The local model quality in contrast, focuses only of the query proteins structure and stability at different points along the amino acid sequence. If the green line is below zero the structure is stable.

Figure 6

Validation of the ReFOLD predicted protein structure using the Ramachandran plot. The Ramachandran plot enables to clearly identify which residues are in an unstable orientation.

Figure 7

Tertiary structure of Cystatin C of *Danio rerio* in Grey coloration with the N-terminal Glycine residue in Red coloration, the QVVAG motif in Green coloration and the PW motif in Blue coloration.

Figure 8

The 2D representation of the protein-ligand complex formed in the active site gorge between AChE Echothiophate. A) Depicts the interactions in hAChE with strong interactions between the ligand and the

characteristic Tryptophan residue. B) Depicts the interaction in rAChE showing tryptophan residues involved in formation of the stable protein-ligand complex.

Figure 9

Structures of Cystatin C (blue) – cysteine protease complexes with clear blockage of the proteases active sites (red). Papain is depicted in grey, Cystatin B depicted in green, Cystatin H depicted in brown, Cystatin L1 depicted in purple and Cystatin S depicted in pink.

Figure 10

Surface view of the protein's tertiary structure (grey) with the active binding site indicated in red.

Figure 11

Human gene interaction network of Cathepsin B (green), Cathepsin L1 (purple) and Cathepsin S (pink). Clinically significant gene interactions are depicted in black.

Supplementary Files

This is a list of supplementary files associated with this preprint. [Click to download.](#)

- [Supplementary.docx](#)