https://doi.org/10.48048/tis.2023.6563

Isolation of Genes Encoding for Human Epidermal Growth Factor (hEGF) from Human Blood: *In Vitro* and *In Silico* Study

Imelda Maelani[1], Suhartono Suhartono[1,*] and Zulkarnain Zulkarnain[2]

[1]Department of Biology, Faculty of Mathematics and Natural Sciences, Syiah Kuala University, Kota Banda Aceh, Aceh 23111, Indonesia
[2]Department of Biomedical Sciences, Faculty of Medicine, Syiah Kuala University, Kota Banda Aceh, Aceh 23111, Indonesia

([*]Corresponding author's e-mail: suhartono@unsyiah.ac.id)

Received: 2 January 2023, Revised: 28 January 2023, Accepted: 3 February 2023, Published: 18 March 2023

Abstract

The 6.12 kDa human epidermal growth factor (hEGF) protein is expressed in the body and plays a role in cell proliferation and differentiation. This research aims to isolate the hEGF genes, determine the phylogenetic relationship between organisms based on the hEGF gene, construct the 3D structure of the hEGF protein, and identify the precise binding position between the hEGF protein and the tested Myosin-9. DNA was isolated from blood of healthy individuals followed by amplification using specified primers. DNA sequencing was utilized to identify the amplification results before generating the phylogenetic trees. The retrieved sequences were then modeled using the Swiss model and docking proteins *in silico.* In three samples, DNA was successfully amplified to 700 bp and the phylogenetic analysis revealed that the three hEGF gene samples belonged to the same clades. With a 30 % identity seq, aligned result sequences modeled using the Swiss model created proteins which were not homologous. The outcomes of this modeling were compared to the hEGF protein from the database, which had a sequence identity of greater than 90 %. The i-RMSD value of the target protein (5.4 +/- 0.4), the Van der Waals energy (-29.3 +/- 1.7), and the Z-score were determined by simulation (-1.7). These results indicate a potential interaction between the isolated EGF protein and the myosin-9 tested.

Keywords: Docking, DNA isolation, hEGF, Phylogenetics, Protein modeling

Introduction

Human epidermal growth factor (hEGF) is a chromosomally encoded protein present in saliva, urine, and blood plasma [1,2]. EGF plays an additional role in signaling pathways and wound healing [3,4]. This growth factor's activity can enhance the proliferation and differentiation of several cells [5]. Thus, one of the applications of hEGF is in the cosmetics industry and clinical sector [6].

Conventional hEGF purification has been performed with relatively poor efficiency. Because it is against recognized bioethics to create this purification directly from an individual, different approaches are required to improve the efficacy of hEGF. Recombinant DNA technology and fusion expression are two methods that can be utilized to create hEGF in big quantities [7]. Various proteins with specific biological roles have been found using in vitro and in silico methods as science and technology have advanced. For the determination of the function of a protein in silico, bioinformatics studies are used with great precision and validation [8].

Several programs can be used to simulate the potential of a molecule through computational molecular analysis using in silico techniques. In silico analyses are particularly helpful for identifying the protein's potential [9]. Docking is a technique used in computational molecular analysis to identify compounds that are geometrically and energetically compatible with the target protein binding sites. This method is beneficial for protein modeling and interactions between the examined protein and the chemicals used to represent it [10]. Employed molecular docking techniques to assess the affinity and interaction between ligands and EGF receptors [11].

On the basis of this research, it has been determined that the ligand molecules utilized have the potential to serve as candidates for anticancer medications that have been simulated using molecular docking. Additionally, protein modeling can be organized according to the sequence of alignment results. The resulting protein model serves as a reference point for further research on the hEGF protein. This study used Myosin-9 because of its antigenicity which can form specific antibodies. It becomes an infector when

interacted with hEGF protein. The goals of this research were to isolate and identify the specific sequence of the gene that codes for the hEGF protein; analyze the evolutionary relationships between different organisms using the hEGF gene; create a three-dimensional model of the hEGF protein, and determine the specific binding location between the target protein (Myosin-9) and hEGF.

Materials and methods

Blood sampling

The sampling procedure was approved by the Health Research Ethics Committee (KEPK) of the Faculty of Medicine, Syiah Kuala University, through an ethical approval letter, 017/EA/FK/2022. The blood sampling procedure was carried out at Dr. Mohd. Zein's Primary Clinic at Universitas Syiah Kuala. A total of three ml blood of healthy individuals with a body weight below 60 kg without history of hypertension and diabetics. The blood of healthy patients was taken using a diagnostic test procedure. Blood was sampled through the median cubital vein using a vacuum tube containing 10 µl EDTA. The collected blood samples were stored in a freezer at -20 ℃ overnight.

DNA extraction

DNA extraction was carried out using the Wizard Kit (Promega, USA) according to the manufacturer's instructions. The DNA extraction included cell lysis, DNA binding, DNA washing, and elution using reagents and buffers. The extracted DNA was subsequently examined quantitatively using a nanophotometer at wavelengths of 260 and 280 nm and qualitatively with gel electrophoresis at 100 V for 40 minutes.

The procedure of characterizing DNA by gel electrophoresis was carried out using a 1 % agarose gel which was preceded by assembly the gel. A whole of 0.65 grams of agarose powder was weighed using an analytical balance. Agarose powder was then mixed with 65 ml of TBE 1x in addition to heated using a microwave intended for 8 minutes awaiting the solution boiled. Furthermore, as much as 0.65 µL red gel was added to the agar mixture. Agarose is poured into the mold and waited for about 10 minutes for it to solidify. After the agarose gel is prepared for use, 1.5 µL of the sample that has been extracted and 0.5 µL of loading dye is put into the well of the gel. This stage is carried out on all samples to be tested. The agarose gel was then electrophoresed at 100 V for 40 minutes. Following the running procedure was successful, afterwards the gel was analyzed by utilizing UV light using a UV transilluminator. The results to be obtained are in the outline of DNA bands with size comparisons referring to the leader DNA.

Amplification of the hEGF protein coding gene

All PCR amplifications were performed in 25 µL reactions containing HotStarTaq master mix 1X 0.4 µM of each primer, namely EGF forward (5′TCCTCTTTGGCAGTCATCCC3') and reverse (5′CATTTCCTGCGAGAGTACCTT3') [12], and 2 µL of template DNA. DEPC-treated water (EMD Millipore, Darmstadt, Germany) was used as no template control (NTC) run in parallel with samples. The PCR reactions were carried out using a PCR thermocycler under conditions as follows: 94˚C for 5 minutes then 35 cycles of the denaturation stage (94˚C for 45 seconds), annealing (57˚C for 45 seconds), initial elongation (72˚C for 45 seconds), and final elongation (72ºC for 5 minutes). The PCR products were analyzed on 1 % (w/v) agarose gels at 100V for 24 min in 0.5X TBE buffer and visualized using GelDoc to assess bands of the expected size.

DNA sequencing and phylogenetic tree construction

DNA sequencing was performed by Genetika Sciences. Ltd. Indonesia using the Sanger method (Sanger dideoxy sequencing). The sequence was further analyzed using Bioedit followed by molecular analysis and phylogenetic tree construction using MEGA version 11. The phylogenetic tree was constructed using sequence data from the five organisms with the highest similarity to the sample sequences. The sequences were then aligned with the software MEGA 11 and the CLUSTAL-W program. The results of the phylogenetic tree construction built in this study referred to the hEGF sequences using the Maximum Parsimony (MP) method with a bootstrap of 1000 replications. The phylogenetic tree construction was carried out by involving 19 sequences, namely three sample sequences, 15 sequences from NCBI, and one outgroup sequence.

***In silico* analysis of hEGF protein**

The potential activity of the protein is predicted *in silico* studies [9]. In this case, a target protein called hEGF was docked using the software to create a 3-dimensional structure called "high ambiguity driven protein-protein docking" (HADDOCK 2.4). This analysis used the principle of binding a three-dimensional structure to the OH group present in the target receptor as the ligand. Furthermore, the bond energy was calculated with the ligand. Three-dimensional production of the hEGF protein was then carried out.

Three-dimensional structure of the hEGF protein

A three-dimensional structure for the protein model was generated using the SWISS-MODEL site. The data consisted of BLAST-processed sequencing results from the database and search results from the Protein Database (PDB) (https://www.rcsb.org). Target protein sequence information was derived from consensus sequencing of each sample and the hEGF protein sequence from the database with the code 1MOX (PDB, 2022). The hEGF protein modeling in this study had 700 bp nucleotides that were translated using the Expassy web into amino acid sequences, yielding predictions for approximately 200 amino acids. The isolated gene positions were ranked between positions 136 and 687 in the hEGF gene database. After obtaining the structure, homology comparisons and analyses were conducted.

Docking

Docking started with the preparation of the protein using the Discovery Studio visualizer. The docking was performed using web servers Computer Atlas of Surface Topography of Protein (CASTp) and HADDOCK 2.4 with specific parameters. The energy involved in this docking was measured by the MolDock Score, also known as the energy value [13].

Myosin-9's structure was downloaded from the PDB web database (scrb.pdb.org). The data were then entered into the program and analyzed using the docking-specific steps. Protein receptor preparation was performed with PDB-tools (cite) utilizing the PDB res application to renumber the PDB file residues beginning at a specific number, followed by docking with HADDOCK 2.4, which took into account the active site of the target protein and the Myosin-9 protein prior to docking. This information was accessible via the CASTp Web server. The process of determining the active site involved uploading and submitting the target protein to the file. Using the Discovery Studio visualizer, the docking was visualized.

Results and discussion

Quantitative analysis of isolated DNA

The isolated DNA quantitatively measured for its concentration and purity was shown in **Table 1**.

Table 1 Results of DNA concentration calculations using a Nanophotometer.

Sample ID	Sample Code	Repeated measurements	DNA consentration (ng/µl)	DNA purity 260/280 nm
1.1	A1	1	21.3	1.68
		2	22.3	1.75
		3	25.4	1.71
1.2	A2	1	14.4	1.92
		2	16.8	1.80
		3	17.0	1.85
1.3	A3	1	12.3	1.72
		2	14.1	1.72
		3	14.0	1.63

Note: Sample IDs (1.1, 1.2, and 1.3) refer to blood samples withdrawn from patient 1, 2, and 3. Sample codes (A1, A2, and A3) refer to DNA extractions using blood samples of patient 1, 2, and 3, respectively.

The results from measuring the purity of DNA for all samples using wavelengths 260 and 280 nm showed values between 1.63 and 1.92. These reference wavelengths are used to assess the quality of extracted DNA [14]. The average value for Sample A2 was optimal at 1.85, which is close to the target ratio of 1.80 for DNA purity according to reference [15]. A value for the OD (optical density) of 260/280 between 1.8 and 2.0 is considered an acceptable range for DNA purity, as noted in reference [16]. However,

the nanophotometer-based OD calculations for samples A1 and A3 showed values lower than 1.8, which may be due to the presence of residual phenol or other substances, as mentioned in [16]. Gel electrophoresis can be used to confirm the quality of the DNA and determine if it is suitable for further testing.

The visualization outcomes of 1 % gel electrophoresis revealed the presence of DNA bands in three high-quality DNA samples (**Figure 1**).

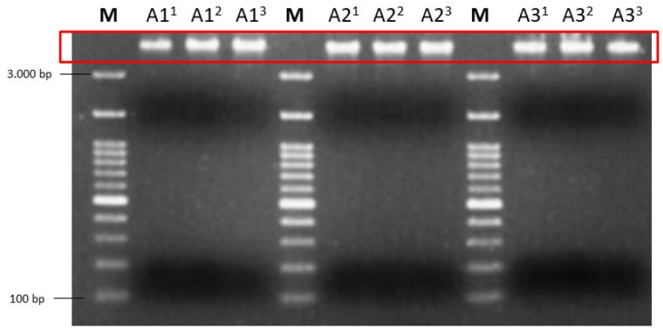

Figure 1 Visualization results of all samples using agarose gel electrophoresis with M: marker (sample A1, $A1^1$: repetition 1, $A1^2$: repetition 2, $A1^3$: repetition 3), M: marker (sample A2, $A2^1$: repetition 1, $A2^2$: repetition 2, $A2^3$: repetition 3), M: marker (sample A3, $A3^1$: repetition 1, $A3^2$: repetition 2, $A3^3$: repetition 3).

The DNA bands in the visualization results are distinct and single-shaped, as shown in Figure 1. This indicates that the process of extracting DNA from human blood samples was successful. High concentrations of DNA would appear as a smear when viewed with gel electrophoresis, but DNA with good quality would show thick, clean bands that are easy to determine under UV light. On the other hand, DNA with low concentration will have thin, hard-to-detect bands. Gel electrophoresis is a method that separates nucleic acids or proteins based on size using an electric current [17]. The quality of the DNA can be determined by examining the appearance of the bands under UV light [18].

Amplification of hEGF protein coding gene

Samples A1, A2, and A3 were amplified at 57°C, which is the temperature recommended for amplifying the hEGF gene [12]. The concentration of EGF primers (10 mM) and the annealing temperature (57°C) were also important factors in the success of the amplification process. The primary attachment, or annealing, is a crucial step in the amplification cycle as it allows the primer to bind to the template and generate the amplicon. The quality of the samples was evaluated using 1 % agarose gel electrophoresis, which was visualized using UV light. The results of the PCR amplification can be seen in **Figure 3**.

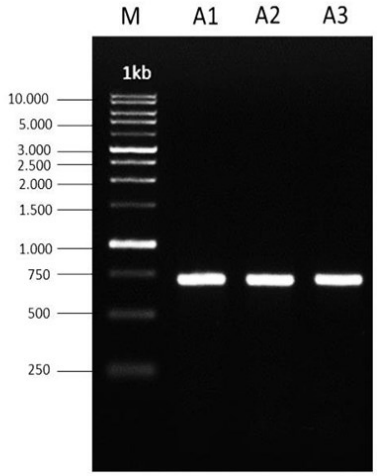

Figure 2 Results of amplification using primers EGF-F and EGF-R with description M: marker, A1: Sample A1, A2: Sample A2, A3: Sample A3.

Figure 2 shows that DNA bands were obtained for samples A1, A2, and A3 using markers of different sizes ranging from 250 bp to 1 kb. All of the samples showed a size of 700 bp, as indicated by a DNA band at the same size as the marker range. The hEGF gene amplification should result in a band of size 708 bp [12]. The samples in this study had strong, pure bands that were not separated, indicating that the hEGF gene amplification was successful and the samples are ready for sequencing. The results of the hEGF gene amplification can be seen in **Figure 2**.

Determination analysis of the sequencing of the hEGF gene

The sequencing results for samples A1, A2, and A3 using primers EGF-F and EGF-R showed peak traces, but the signal was lost at the beginning and end of the sequence, resulting in a buildup of peaks. This could be caused by salt contamination in the peak trace. Adequate sequencing results should have clear, distinct, and separate peak traces [19]. The nucleotide sequences obtained from the sequencing were analyzed using Bioedit software and a consensus was formed by combining all of the sequences representing each amplicon [20]. The consensus was then submitted to the NCBI website for BLAST-N analysis, the results of which can be found in **Table 2**.

Table 2 Result of BLAST-N analysis.

Description	Sample identity	Query Cover	E. Value	Percent Identity	Accession
Homo sapiens	A1	100 %	0.0	99.85 %	AY506357.1
Homo sapiens		100 %	0.0	99.71 %	NG_011441.2
Homo sapiens		100 %	0.0	99.71 %	AC005509.1
Chlorocebus sabaeus		99 %	0.0	99.51 %	XM_037990891.1
Chlorocebus sabaeus		99 %	0.0	99.51 %	XM_007999539.2
Homo sapiens	A2	100 %	0.0	99.56 %	NG_011441.2
Homo sapiens		100 %	0.0	99.56 %	AY506357.1
Homo sapiens		100 %	0.0	99.56 %	AC005509.1
Chlorocebus sabaeus		99 %	0.0	99.23 %	XM_037990891.1
Chlorocebus sabaeus		99 %	0.0	99.23 %	XM_007999539.1
Homo sapiens	A3	100 %	0.0	99.56 %	NG_011441.2
Homo sapiens		100 %	0.0	99.56 %	AY506357.1
Homo sapiens		100 %	0.0	99.56 %	AC005509.1
Chlorocebus sabaeus		99 %	0.0	99.21 %	XM_037990891.1
Chlorocebus sabaeus		99 %	0.0	99.21 %	XM_0079999539.2

Table 2 shows that samples A1, A2, and A3 have similar sequences to several organisms. The top five sequences are the alignments of samples A1, A2, and A3 with the highest similarity to the query, with a coverage of 100 %. Sample A1 had a similarity of 99.85 %, sample A2 had 99.56 %, and sample A3 had 99.56 %. The results of the comparison with the five organisms were saved in Fasta format and used to construct a phylogenetic tree.

Construction analysis of phylogenetic trees

In this study, *Tarsius dentatus* was used as an outgroup in building the phylogenetic tree. Outgroups are chosen based on the relationship of the sample being analyzed [21]. *Tarsius* is a type of primate known for being small, nocturnal, and having large eyes [22]. Tarsius can be separated into different clades of *Homo sapiens*, which makes it a suitable outgroup for constructing a phylogenetic tree [23]. The results of the tree construction can be seen in **Figure 4**.

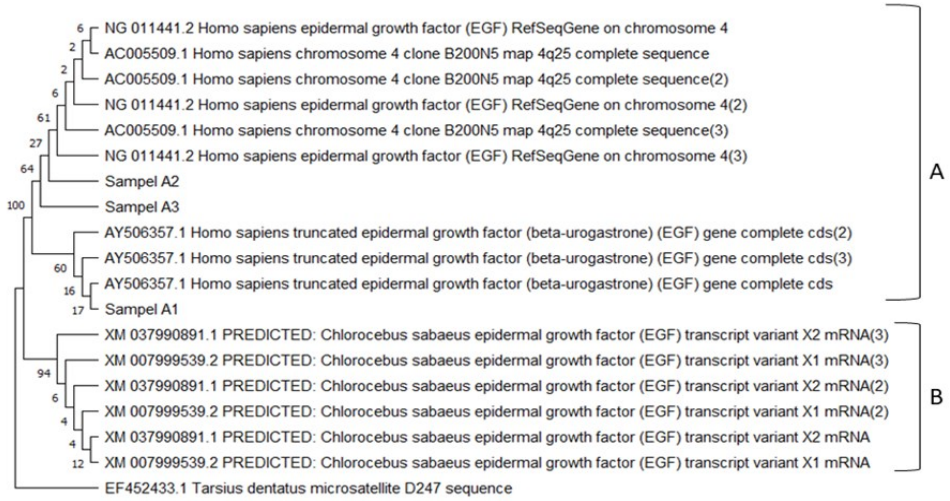

Figure 3 The results of constructing a phylogenetic tree based on hEGF sequences using the MP method with a bootstrap 1000 times.

Figure 3 shows the results of building a phylogenetic tree for samples A1, A2, and A3, which are closely related. The tree shows the formation of groups, or clades, between the database sequences and the test samples. There are two main groups in the tree: Group A with a bootstrap value of 100 % and Group B with a bootstrap of 94 %. Group A has two branches, with bootstrap values of 60 and 64 %. Samples A2 and A3 are closely related to the hEGF gene from *Homo sapiens*, with a bootstrap value of 64 %. Sample A1 has a bootstrap value of 60 % and is also closely related to the hEGF gene from *Homo sapiens*. The phylogenetic tree forms a monophyletic group, which is a group of taxa that share a common ancestor [24]

In the phylogenetic tree, the hEGF gene from *Chlorocebus sabaeus* (the green monkey) dominates Clade B with a bootstrap value of 94 %. This suggests that the samples are not closely related. The tree shows that Clade B meaning that it is made up of multiple ancestral lineages, while *Chlorocebus sabaeus* is divided into clades but is on the same branch [25]

3-dimensional structure analysis

Identification template

The 3D structure obtained was based on the search for the template. The results of the mold model selection are determined by several parameters, such as the global model quality estimation (GMQE), identity, oligo site, and ligands. Based on the print search results on the Swiss model server, data were obtained as shown in **Table 3**.

Table 3 Modeling parameters using the Swiss model.

Parameters	Protein samples			
	A1	**A2**	**A3**	**1MOX**
Template	5clb.1A	2l3x.1.A	2ygj.i.A	7sz5.1
Identity	27.7 %	28.57 %	22.50 %	99.89 %
Oligo site	Monomer	Monomer	Monomer	Hetero-tetramer

Model evaluation

A template serves as a model for each protein sample. The search results showed that sample A1 had 7 mold models, sample A2 had 14 models, sample A3 had 15 models, and 1MOX had 50 models. The structure with the highest identity to the original form was selected. According to **Table 3**, the identity for the isolated protein samples (A1, A2, and A3) was highest for sample A2, with a value of 28.57 %, while the identity with code 1MOX was 99.82 %. This value reflects the diversity of amino acid residues in the

target protein sequence [26]. Multiple variations of each mold model may be obtained based on alignment homology, quality, and structural flexibility, which can affect the accuracy of the modeling results [27].

However, the identity for sample A2 was below 30 %, which means that the amino acid sequence of the target protein had a low level of similarity to the template used. When compared to samples A1, A2, and A3, 1MOX had an identity above 90 %. This value represents the degree of similarity between the target protein and the selected template model that can be aligned. A percentage of similarity below 25 % indicates a significant difference between the target protein and the selected template [28]. The results of 3D modeling using the Swiss model can be seen in **Figure 4**.

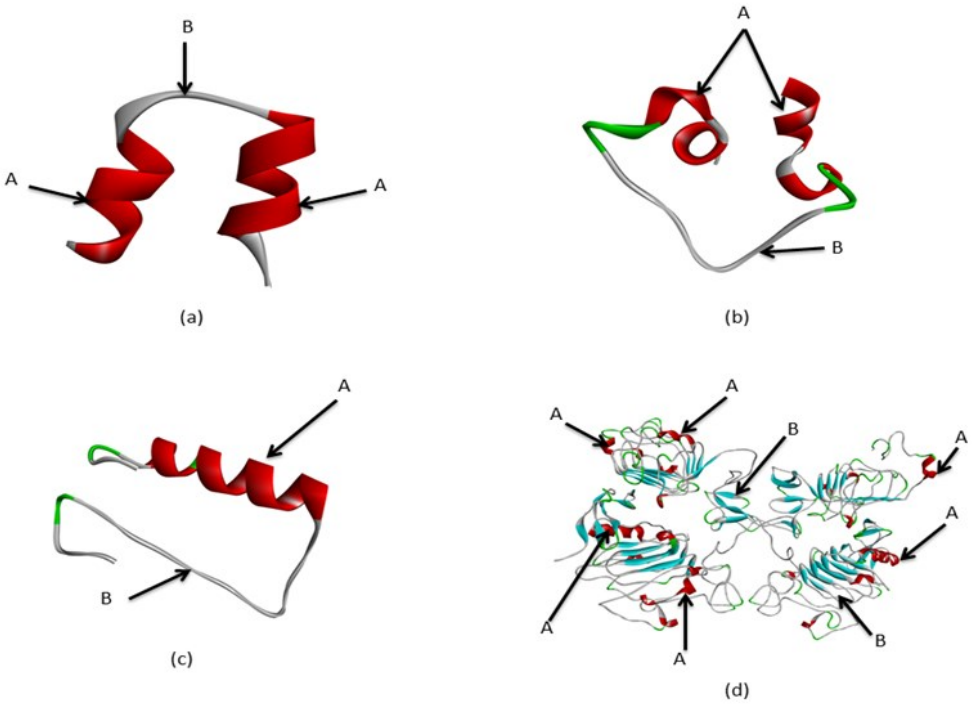

Figure 4 Results of 3D modeling using the Swiss model with A: alpha-helix, B: Beta-sheet. (a) sample A1, (b) sample A2, (c) sample A3, and (d) 1MOX.

Figure 4 shows that samples A1, A2, and A3 have multiple alpha-helix and beta-sheet regions, while sample 1MOX has more of these structures than the other samples. The amino acid composition of these structures can affect their formation. According to the structure, the amino acid side chains determine the unique shapes of the alpha-helix and beta-sheet [29]. These two structures have different compositions. The alpha-helix structure is a spiral arrangement of carboxyl and amine groups that are bonded by hydrogen bonds, resulting in a spiral shape. On the other hand, the beta-sheet structure is formed by reverse folding on each side.

The three-dimensional modeling of the hEGF protein in each of the obtained samples reveals a secondary structure based on the 3D formation of the alpha-helix and beta-sheet (**Figure 4**). In contrast, the hEGF protein with code 1MOX has a more complex structure with up to four times more complexity. This significant difference is due to the presence of multiple stop codons in the Expassy web translation results, which disrupt the open reading frame (ORF) and affect the translation results. As a result, the protein structure generated in this study is shorter than the protein stored in the protein database.

Based on the highest identity score, sample A2 was further evaluated to determine its suitability as a representative of the isolated sample. In subsequent evaluations, the hEGF protein structure with access code 1MOX was used as a comparison in the database. The Ramachandran plots and MolProbity result can be used to evaluate the protein structure formed on servers using specific values and graphs. These two parameters indicate the quality of a protein model. **Figure 5** shows the results of the Ramachandran plots for samples A2 and 1MOX.

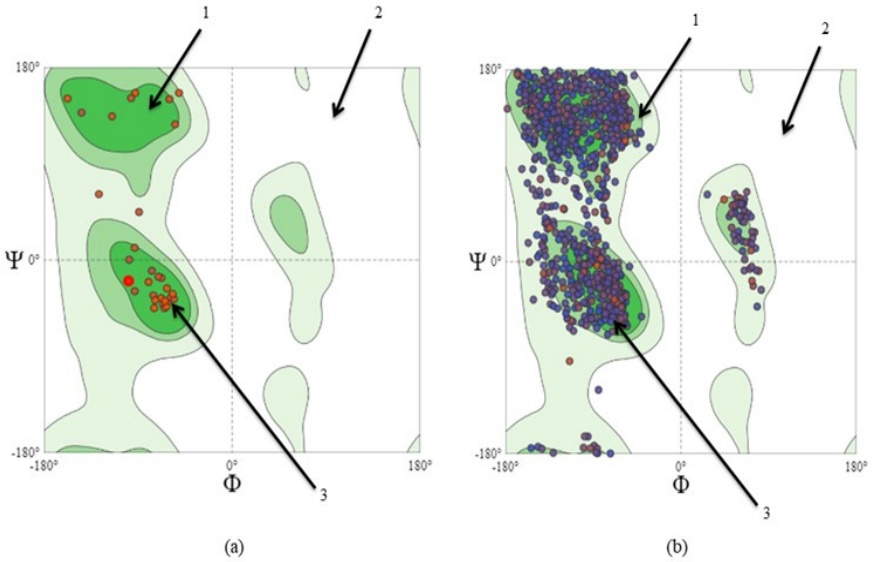

Figure 5 Ramachandran plots on (a) model A2, (b) 1MOX model with 1: Ramachandran favored, 2: Ramachandran outlier, 3: Amino acid residue.

Based on the Ramachandran plots, the protein has a relatively good quality in terms of the overall number of amino acids. There are more amino acids in the favored region than in the outliers. **Figure 5(a)** shows the distribution of the amino acids in the A2 model, with more amino acid residues in the favored region than in the outliers. Similarly, the 1MOX model (**Figure 5(b)**) is more dominant in the favored area. These results meet the quality standards for 3D modeling. More detailed information about the results of the A2 model analysis can be found in **Table 4**.

Table 4 Modeling result of A2 sample.

Parameters	Analysis results
MolProbity score	2.15
Clash score	1.82
Ramachandran favored	87.88 %
Ramachandran outliers	3.03 % (A145 PRO)
Rotamer outliers	6.25 % (A141 LYS, A153 SER)
C-beta deviations	1 (A153 SER)
Bad bonds	0/283
Bad angles	4/379 (A128 HIS, A154 HIS, A126 HIS, A134 HIS)

Evaluation of the modeling results by comparing the Quality Model Energy Analysis (QMEAN) and GMQE values showed that the QMEAN values for samples A2 and 1MOX were 0.43 ± 0.12 and 0.77 ± 0.05, respectively, while the GMQE values for samples A2 and 1MOX were 0.05 and 0.88 (Table 3). These values meet the modeling standards, with the GMQE value being close to 1 [26]. The QMEAN and GMQE values provide estimates of the quality of a model by combining several protein functions. The GMQE value indicates the suitability between the target protein structure and the template model [29]. The quality of the modeling results is also supported by the favored and outlier data (**Table 4**). The modeling results are based on several parameters, such as the MolProbity Score, Clash Score, Ramachandran Favored, and others. The results of the subsequent analysis of the 1MOX model can be seen in **Table 5**.

Table 5 Modeling result of 1MOX.

Parameters	Analysis result
MolProbity Score	1.99
Clash Score	2.27 (C397 GLU-C427 ARG), (C118 GLU-C198 ARG)
Ramachandran Favored	86.46 %
Ramachandran Outliers	0.73 % A432 ILE, C327 ILE, C112 PRO, A128 ASN, A136 GLU, A145 SER, C441 GLY, A207 CYS
Rotamer Outliers	2.90 % A98 LEU, A195 CYS, A423 SER, C207 CYS, A149 LEU, A70 ASN, A138 ILE, A462 GLN, C38 LEU, D1 VAL, A147 ASP, B49 LEU, A368 LEU, C431 GLU, C437 VAL, C419 LEU, B1 VAL, A136 GLU, B2 VAL, C108 LEU, A194 GLN, C232 ASP, A144 VAL, C498 ASP, B33 VAL, C420 ASN, C109 LYS, C156 PHE
C-Beta Deviations	20 A418 SER, C418 SER, A468 SER, A91 ASN, A99 SER, C339 THR, C468 SER, B10 ASP, C459 THR, C378 THR, C297 ASP, C174 SER, D10 ASP, A297 ASP, B50 ALA, C154 MET, B6 ASN, A310 ARG, C373 THR, C156 PHE
Bad Bonds	12 / 8730 A121 HIS, C334 HIS, A280 HIS, A359 HIS, C409 HIS, C359 HIS, B45 HIS, B18 HIS, A483 HIS, A409 HIS, D12 HIS, C394 HIS
Bad Angles	107 / 11826 A355 ASP, C232 ASP, C344 ASP, C154 MET, A279 ASP, A232 ASP, C182 ASN, C280 HIS, A40 ASN, C373 THR, (C240 CYS-C241 PRO), (B8 CYS-B9 PRO), C290 ASP, C279 ASP, C156 PHE, B47 ASP, (A428 SER-A429 LEU), A403 ARG, (C133 CYS-C134 ASN), A131 ALA, (A307 GLY-A308 PRO), A238 ASP, C51 ASP, C134 ASN, C339 THR, A280 HIS, C355 ASP, B4 HIS, D47 ASP, C412 PHE, (A1 LEU-A2 GLU), B10 ASP, (A247 ASN-A248 PRO), A121 HIS, A147 ASP, A392 ASP, A346 HIS, (A386 TRP-A387 PRO), A70 ASN, A442 ASN, C420 ASN, (A278 THR-A279 ASP), D12 HIS, (A241 PRO-A242 PRO), (A495 GLU-A496 PRO), (A75 ILE-A76 PRO), C366 GLN, B45 HIS, C392 ASP, (D8 CYS-D9 PRO), A23 HIS, (A479 GLY-A480 GLN), B18 HIS, D10 ASP, B7 ASP, C23 HIS, A483 HIS, B6 ASN, C305 CYS, C359 HIS, (C307 GLY-C308 PRO), C152 MET, C394 HIS, (C256 ASN-C257 PRO), A334 HIS, C209 HIS, D4 HIS, C159 HIS, C238 ASP, A209 HIS, C121 HIS, C409 HIS, D18 HIS, (A487 SER-A488 PRO), A187 THR, A409 HIS, D45 HIS, D35 HIS, (C218 GLY-C219 PRO), A359 HIS, A364 ASP, (A364 ASP-A365 PRO), D6 ASN, C334 HIS, C459 THR, (C364 ASP-C365 PRO), (A127 SER-A128 ASN), (A397 GLU-A398 ASN), A459 THR, A206 ASP

Outlier regions are one of the benchmarks for determining the quality of a protein structure. If we look at the results of the Ramachandran outlier analysis for samples A2 and 1MOX (**Tables 4** and **5**), we see that they have values of 3.03 and 0.73 %, respectively. This score indicates that the quality of these protein models is not yet good. The quality of a protein structure is considered good if the number of amino acid residues in the outlier is below 0.2 % [30]. On the other hand, based on the number of amino acid residues in the favored and outlier regions, the favored Ramachandran value for sample A2 (**Table 4**) was 87.88 %, while the outlier Ramachandran was 3.03 %. These results indicate that the protein model structure is quite good. The results in Table 5 showed a favored Ramachandran of 86.46 % and an outlier Ramachandran of 0.73 %. These results also indicate that the protein model structure in the 1MOX modeling is quite good since the amino acid residues are more dominant in the favored region than in the outlier region [26].

Another parameter that can be observed is the MolProbity value. The MolProbity obtained from the A2 model of this study was 2.15, while 1MOX model was 1.99. The clash score and the total Ramachandran percentage rotamer are combined to produce this value. The MolProbity must be below the true crystallographic resolution [31]. Based on this description, it is suspected that the 3-dimensional structure of samples A1, A2, and A3 is not homologous to the existing structure, while the 1MOX structure is homologous to the existing protein structure in the database.

Analysis of docking

The docking analysis in this study consisted of three main stages that took several hours. The first stage of this process is determining the binding position (rigid body docking). This information is useful for estimating the interfacial interactions between proteins. The obtained results went through a process of improving the rotation angle (torsion angle) of the protein called it1. The results of this study include three simulations of improvements to the annealing process, which are carried out in stages. The annealing process consists of 500 stages with an optimized rigid body. The annealing process includes 1000 steps during which the side chain on the interface is switched. The annealing process includes 1000 steps during which both the side chain and the back side of the interface are switched to form a new conformation. The results of the it1 stage go through the final repair process (itw) to reduce the energy level [32].

The protein sequences analyzed in this study were the modeling results that had the most optimal values among the other samples. The samples used for the docking were A2 and 1MOX; each of these proteins went through the docking with the same Myosin-9. The myosin-9 used in this study is myosin, derived from a database with access code 4CFQ (https://www.rcsb.org) [33]. The number of amino acid residues in each target protein varied. The A2 protein consisted of 35 amino acid residues, while 1MOX consisted of 1,101 amino acid residues. In addition, the docking on the HADDOCK web has been sorted based on the best value, known as the cluster. Each cluster contained several parameters that can be used to predict the target protein's interaction with Myosin-9. Evaluation of docking was based on the HADDOCK score parameter, the root mean square deviation (RMSD) value, the Van der Waals energy, and the z-score The RMSD value is a value that indicates how much transformation occurs between the target protein and the molecule being tested. The quality of the docking is classified as valid if the RMSD value is 2 [34].

Based on the docking of the target protein (A2) with Myosin-9, HADDOCK groups 148 structures. A total of identified structures were grouped into 10 clusters, where 74 % represents the amount of water refined (water molecules) that has been purified. There were 200 models have been successfully clustered on Hadoop 2.4 servers. The number of models was then categorized into 10 clusters with the best value. Figure 6 describes a comparison of HADDOCK values with the RMSD interface (i-RMSD) between EGF protein (A2) and Myosin-9. The results obtained show that the most stable interaction is in cluster 3, with a HADDOCK score of -70.3 +/- 8.6 and an i-RMSD of 2.7 +/- 1.6. Based on this value, the form of the model and the interactions formed in cluster 3 were considered to have met the modeling standards.

A simulation was performed to examine the interaction between the docking target protein (1MOX) and Myosin-9. The simulation identified 128 structures, 64 % of which were water-refined and purified. The structures were then grouped into 9 clusters using the HADDOCK server. The results indicated that the most stable interaction occurred in cluster 3, with a HADDOCK score of -66.3 +/- 3.0 and an i-RMSD of 5.4 +/- 0.4. This modeling in cluster 3 met the standard for acceptable protein-protein docking, as defined by an i-RMSD value of 2 or less.

The quality of the docking simulation can be evaluated by examining the standard deviation of the data, which can be represented by the z-score. A lower z-score indicates better modeling results for protein docking. In the docking simulation using HADDOCK 2.4, the interaction between the EGF protein (A2) and Myosin-9 resulted in a z-score of -1.3 in cluster 3, while the interaction between EGF (1 MOX) and Myosin-9 resulted in a z-score of -1.7 in cluster 3. Both of these results are considered valid and can be used in further testing, as shown in **Figure 6**.

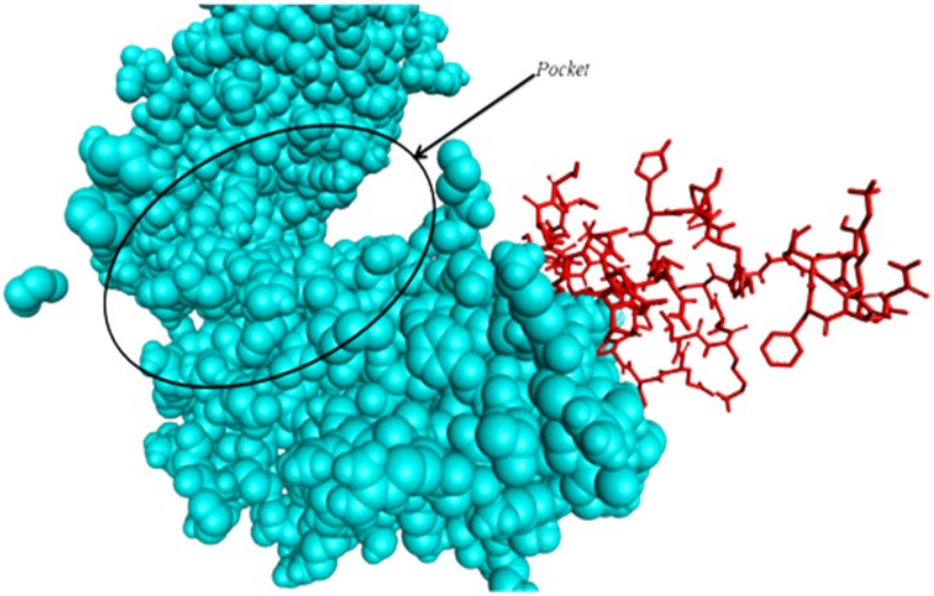

Figure 6 Binding position of EGF protein (A2) with Miosin-9, red color; EGF in stick; blue color: Miosin-9 in surface.

The active site of the EGF protein (A2) was found to bind to pocket 1 using the CASTp web server, with an area of 17,356 and a volume of 5,530. The binding position between the EGF protein (A2) and Myosin-9, as shown in **Figure 6**, indicates that the interaction occurs on the active site of Myosin-9. The topography of the EGF (1 MOX) protein also revealed an active site with an area of 4,533,742 and a volume of 6,183,525. The three-dimensional modeling of the intermolecular interactions is depicted in **Figure 7**.

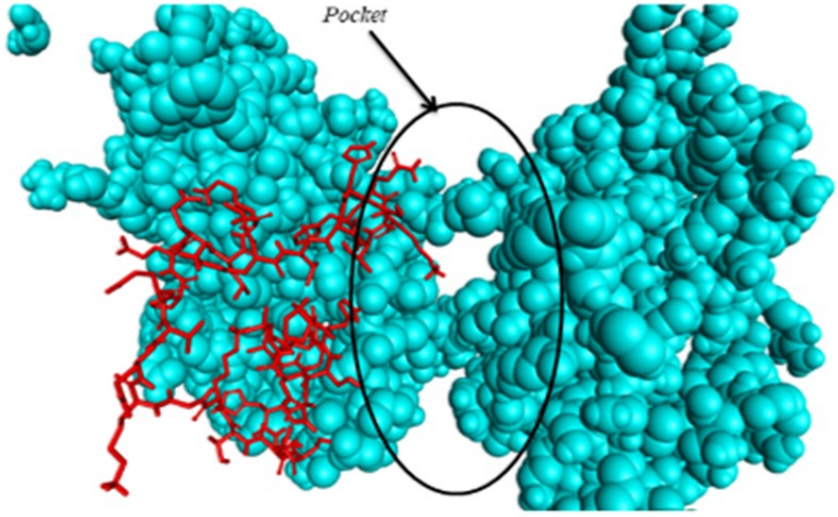

Figure 7 The binding position of EGF protein (1MOX) with Miosin-9, red color; EGF in stick; blue color: Miosin-9 in surface.

Figure 7 shows the interactions that occur in the pocket of Myosin-9 as a result of the binding position of the EGF (1 MOX) protein with Myosin-9. Both models produced in the docking simulation met the necessary standards and can be used in further testing to explore the interaction between the two proteins.

Conclusions

The findings of this study indicate that the gene encoding hEGF is amplified to a size of 700 bp, and samples A1, A2, and A3 form a monophyletic group in the phylogenetic tree. Protein modeling identified the identical sequences of samples A1 (27.7 %), A2 (28.57 %), A3 (22.50 %), and 1MOX (99.89 %). The i-RMSD values from docking sample A2 with Myosin-9 and 1MOX with Myosin-9 were 1.6 and 0.4, respectively, indicating that both samples interact with Myosin-9.

Acknowledgements

The authors acknowledged the Genetics and Molecular Biology Laboratory, Department of Biology, Syiah Kuala University, Indonesia for facilitating the study.

References

[1] K Vinay, A K Abbas, N Fausto and JC Aster. *Robbins and Cotran pathologic basis of disease, professional edition e-book*. Elsevier health sciences, Canada, 2014, p. 338-341.
[2] *Epidermal Growth Factor*, available at: https://www.ncbi.nlm.nih.gov, accessed December 2022.
[3] Y Kang, W He, C Ren, J Qiao, Q Guo, J Hu and L Wang. Advances in targeted therapy mainly based on signal pathways for nasopharyngeal carcinoma. *Signal Transduct. Targeted Ther.* 2020; **5**, 1-20.
[4] J Zhu, G Jiang, W Hong, Y Zhang, B Xu, G Song and L Ruan. Rapid gelation of oxidized hyaluronic acid and succinyl chitosan for integration with insulin-loaded micelles and epidermal growth factor on diabetic wound healing. *Mater. Sci. Eng.* 2020; **117**, 1-12.
[5] FD Crendhuty and S Megantara. Sediaan Hydrogel mengandung Epidermal Growth Factor dalam Penyembuhan Luka. *Farmaka* 2019; **17**, 410-6.
[6] S Pouranvari, F Ebrahimi, G Javadi and B Maddah. Cloning, expression, and cost effective purification of authentic human epidermal growth factor with high activity. *Iranian Red Crescent Med. J.* 2016; **18**, 1-8.
[7] X Zheng, X Wu, X Fu, D Dai and F Wang. Expression and purification of human *Epidermal Growth Factor* (hEGF) fused with GB1. *Biotechnol. Biotechnol. Equip.* 2016; **30**, 813-8.
[8] GA Sandag and SW Taju. Bioinformatics tools for data processing and prediction of protein function. *CogITo Smart J*. 2018; **4**, 305-15.
[9] R Solfaine, L Muniroh and WW Mubarokah. Study of doking moleculer flavonoid Coleus amboinicus in TGF-1b receptor and lowering MDA concentration on Cisplatin-induce Wistar Rats. *Jurnal Sains Veteriner* 2020; **38**, 151-8.
[10] R Prasetiawati, M Suherman, B Permana and R Rahmawati. Molecular docking study of anthocyanidin compounds against Epidermal Growth Factor Receptor (EGFR) as anti-lung cancer. *Indonesian J. Pharmaceut. Sci. Tech.* 2021; **8**, 8-20.
[11] S Suherlan, R Rohayah and TM Fakih. Uji aktivias antikanker payudara senyawa andrografolida dari tumbuhan sambiloto (Andrographis paniculata (Burm F) Ness.) terhadap human epidermal growth factor receptor 2 (HER-2) secara in silico. *Jurnal Ilmiah Farmasi Farmasyifa* 2021, **4**, 39-50.
[12] M Malek, F Mashayekhi and Z Salehi. Epidermal growth factor +61A/G (rs4444903) promoter polymorphism and serum levels are linked to idiopathic male infertility. *British J. Biomed. Sci.* 2020; **78**, 92-4.
[13] S Siswandono. *Pengembangan Obat Baru*, Cetakan I. Airlangga University Press, Surabaya, 2014, p. 59-60.
[14] T Pervaiz, X Sun, Y Zhang, R Tao, J Zhang and J Fang. Association between chloroplast and mitochondrial DNA sequences in Chinese Prunus genotypes (*Prunus persica*, *Prunus domestica*, and *Prunus avium*). *BMC Plant Biol*. 2015; **15**, 1-10.
[15] G Koetsir and C Eric. A practical guide to analyzing nucleic acid concentration and purify with microvolume spectrophotometers. *New England BioLabs* 2019; **7**, 1-9.
[16] M Ghaheri, D Kahrizi, K Yari, A Babaie, RS Suthar and E Kazemi. A comparative evaluation of four DNA extraction protocols from whole blood sample. *Cellular Mol. Biol.* 2016; **62**, 120-4.
[17] M Mitra. DNA sequencing basics and its applications. *SCIOL Genet Sci*. 2018; **1**, 80-4.
[18] DW Ardiana. Teknik isolasi DNA genom tanaman pepaya dan jeruk dengan menggunakan modifikasi bufer CTAB. *Buletin Teknik Pertanian* 2009; **14**, 12-6.
[19] Customized Products. Available at: https://base-asia.com/wp-content/uploads/2021, accessed November 2022.

[20] W Windasari, MF Islam and A Taher. Analisis Keragaman Genetik Daerah 3'utr Gen Reseptor Lipoprotein Densitas Rendah (LDLR) Dari Individu Asal Papua. *In:* Prosiding Seminar Nasional MIPA UNIPA, Universitas Papua, Manokwari. 2022, p. 157-63.
[21] JG Rohwer. Toward a phylogenetic classification of the Lauraceae: Evidence from matK sequences. *Syst. Bot.* 2000; **25**, 60-71.
[22] R Widayanti and S Trini. Studi keragaman genetik *Tarsius* sp. asal kalimantan, sumatera, dan sulawesi berdasarkan sekuen gen NADH dehidrogenase sub-unit 4L (ND4L). *Jurnal Kedokteran Hewan.* 2012; **6**, 105-11.
[23] R Widayanti, S Trini and TA Wayan. Keragaman genetik gen NADH dehydrogenase subunit 6 pada monyet hantu (*Tarsius* Sp.). *Jurnal Veteriner*. 2013; **14**, 239-49.
[24] AS Darupamenang, JK Beivy, SA Nio and ET Trina. Analisis Filigenetik Genus Alocasia. *Jurnal Bios Logos* 2022; **12**, 157-63.
[25] AR Simbolon, M Ompi, E Widyastuti and DA Wulandari. Penggunaan DNA barcoding dalam mengidentifikasi larva gastropoda (family cymatiidae) di perairan kepulauan sangihe-talaud, sulawesi utara. *Bawal Widya Riset Perikanan Tangkap.* 2021; **13**, 145-55.
[26] N Komari, S Hadi and E Suhartono. Pemodelan Protein dengan Homology Modeling menggunakan SWISS-MODEL: Protein Modeling with Homology Modeling using SWISS-MODEL. *Jurnal Jejaring Matematika dan Sains* 2020; **2**, 65-7.
[27] F Kiefer, K Arnold, M Künzli, L Bordoli and T Schwede. The swiss-model repository and associated resources. *Nucleic Acids Res.* 2009; **37**, 387-92.
[28] FZ Saudale. Pemodelan Homologi komparatif struktur 3d protein dalam desain dan pengembangan obat. *Al-Kimia* 2020; **8**, 1-11.
[29] S Suprianto and IM Budiarsa. Analisis in silico protein clock (circadian locomotor output cycles kaput) pada Patagioenas fasciata monilis. *Jurnal Sains dan Teknologi* 2021; **10**, 9-15.
[30] Y Yeni and DH Tjahjono. Homology modeling epitop isocitrate dehydrogenase tipe 1 (r132h) menggunakan modeller, i-tasser dan (Ps) 2 untuk vaksin glioma. *Farmasains: Jurnal Ilmiah Ilmu Kefarmasian* 2017; **4**, 21-32.
[31] R Chen. Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnol. Adv.* 2012; **30**. 1102.
[32] V Charitou, SCV Keulen and AM Bonvin. Cyclization and docking protocol for cyclic peptide-protein modeling using HADDOCK2.4. *J. Chem. Theor. Comput.* 2022; **18**, 4027-40.
[33] A Duelli, B Kiss, I Lundholm, A Bodor, MV Petoukhov, DI Svergun, N Latzio and G Katona. The C-terminal random coil region tunes the Ca2+-binding affinity of S100A4 through conformational activation. *PloS One* 2014; **9**, 97654.
[34] FZ Muttaqin, H Ismail and Muhammad. Studi molecular docking, molecular dynamic, dan prediksi toksisitas senyawa turunan alkaloid naftiridin sebagai inhibitor protein kasein kinase 2-A pada kanker leukemia. *J. Pharmacoscript* 2019; **2**. 49-64.
[35] MF Lensink, N Nadzirin, S Velankar and SJ Wodak. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins Struct. Funct. Bioinform.* 2020; **88**, 916-38.