Open Journal of Genetics, 2013, 3, 183-194 **OJGen**
http://dx.doi.org/10.4236/ojgen.2013.33021 Published Online September 2013 (http://www.scirp.org/journal/ojgen/)

Correctness and accuracy of template-based modeled single chain fragment variable (scFv) protein anti-breast cancer cell line (MCF-7)

Elham O. Mahgoub, Ahmed Bolad

Alneelain Medical Research Center, Faculty of Medicine Department of Microbiology and Unit of Immunology, Al-Neelain University, Khartoum, Sudan
Email: ilhamomer@yahoo.com

Received 8 February 2013; revised 20 March 2013; accepted 3 April 2013

ABSTRACT

Multiple sequence alignments can be used in the template-based modelling of protein structures to build fragment-based assembly models. Therefore, useful functional information on the 3D structure of the anti-MCF-7 scFv protein can be obtained using available bioinformatics tools. This paper utilises several commonly-used bioinformatics tools and databases, including BLAST (Basic Local Alignment Search Tool), GenBank, PDB (Protein Data Bank), KABAT numbering and SWISS-MODEL, to gain specific functional insights into the anti-MCF-7 scFv protein and the assembly of single-chain fragment variable (scFv) antibodies, which consist of a variable heavy chain (VH) and a variable light chain (VL) connected by the linker $(Gly_4\text{-}Ser)_3$. The linker has been built as a loop structure using the Insight II software. The accuracy of the loop structure has been evaluated using Root Mean Square Deviation (RMSD). The accuracies of the VL and VH template-based structures are enhanced by using the evaluation methods Verify3D, ERRAT and Ramchandran plotting, which measure the error in the residues. In the results, 100% of the light-chain residues scored above 0.2, whereas 88.5% of the heavy-chain residues' scored above 0.15 in the Verify3D evaluation method. Meanwhile, using ERRAT, the alignments of both chains scored more than 70% in space. Additionally, the Ramchandran plot evaluation method showed large numbers of residues in the favoured areas in both chains; these findings demonstrated that all of the chosen templates were the best candidates.

Keywords: Single Chain Fragment Variable; Homology Modeling; SWISS-MODEL; Insight II; Model Evaluation Method

1. INTRODUCTION

The prediction of protein structure is one of the most important goals pursued by bioinformatics and theoretical chemistry. The scFv anti-MCF-7 gene was constructed from the mouse B-cell hybridoma line C3A8 using phage display technology in a previous study. The objective of scFv protein homology modelling is to predict the three-dimensional structure of the VH and VL chains of the scFv protein from their amino acid sequences. Modelling prediction includes additional relevant information, such as the structures of related proteins. In other words, it deals with the prediction of a protein's tertiary structure from its primary structure. Chua *et al.* [1] investigated many uses for this technology in scFv (single-chain variable fragment) genes cloned from anti-CMV (anti-cucumber mosaic virus). The scFv anti-MCF-7 antibody structure is modelled using SWISS-MODEL, and the VH and VL models are connected by the linker $(Gly_4\text{-}Ser)_3$ in the Insight II software. Thus, the complimentary-determining regions CDRs in the modelled antibody structure are determined by KABAT numbering and mapped to provide insight for further epitopes analysis.

Homology modelling is based on the reasonable theory that two homologous proteins will share very similar structures. Because a protein's folding is more evolutionarily conserved than its amino acid sequence, a target sequence can be modelled with reasonable accuracy on a very distantly related template, provided that the relationship between the target and the template can be discerned through sequence alignment. Homology modelling

was first applied by Tom Blundell in the late 1970's, using early computer imaging methods [2]. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction, given a known-good alignment [3]. Unsurprisingly, suggested homology modelling is most accurate when the target and template have similar sequences. Modeller is a popular software tool for producing homology models using methodology derived from NMR spectroscopy data processing.

The standard procedure of template-based modelling consists of four steps: 1) finding known structures (templates) related to the sequence to be modelled (target); 2) aligning the target sequence onto the template structures; 3) building the structural framework by copying the aligned regions, or by satisfying spatial constraints from the templates; 4) constructing the unaligned loop regions and adding side-chain atoms. The first two steps are usually performed as a single procedure because the correct selection of templates relies on their accurate alignment with the target [4]. Similarly, the last two steps are also performed simultaneously because the atoms of the core and loop regions interact closely. SWISS-MODEL provides an automated web server for basic homology modelling. Accordingly, models are pre-computed similarity relationships between sequences, structures and binding sites [5].

Structure evaluation is the most important component of structure prediction. There are several methods to evaluate protein structures, such as Ramchandran plotting, Verify3D and ERRAT. These programmes are freely available at the UCLA-DOE server. Moreover, the Ramchandran plot was developed by Gopalsamudran Narayana [6], and Verify3D was demonstrated by Eisenberg [7]. In this study, the heavy and light chains are modelled using SWISS-MODEL and connected together with the peptide linker $(Gly4\text{-}Ser)_3$, which was built using the Insight II software. The CDRs in the modelled antibody structure were determined by KABAT numbering and mapped inside the model structures. Moreover, the template structures of the heavy and light chains were evaluated to gain confidence about the correctness of the predicted structures.

2. METHODS

2.1. Protein Homology Modelling of the Heavy Chain and Light Chain

All of the procedures were performed to predict protein structure through homology modelling. First, the ExPASy website (http://www.us.expasy.org/tools/dna.html) was used to translate the nucleotide sequence into the protein sequence. Next, the amino acid sequences of the VH and VL chains were submitted to ncbi-genbank (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the template structures with the highest percentage of alignment. Additionally, similarity was confirmed between the VH and VL sequences and their template sequences. The alignment between the sequences was refined manually using Pairwise (http://www/search/pairwise.shtml). The alignment was obtained from the Pairwise website, and then Cluster-X software was used to predict the VH and VL protein models. The alignment between the sequences was then submitted to the SWISS-MODEL Automated Comparative Protein Modelling Server website (http://swissmodel.expasy.org/workspace/index).

The structure was visualised with the Accelrys Visualize software (http://www.accelrys.com). Also, the models were represented as ribbons generated using the Discover software from Accelrys (San Diego, CA, USA) The higher sequence similarity of the combining sites of VH and VL of the scFv protein was then used to construct 3D structures. Furthermore, comparing the amino acids against the DNA was allowed to construct realistic models of the VH and VL chains of the scFv protein. The target amino acids were manually changed until they were similar to the 3BKY and 1AY1 sequences.

2.2. Build scFv Full Structure Using Builder/Insight Ii Software

The Builder/Insight II software was used to connect the VH and VL models using the linker $(Gly_4\text{-}Ser)_3$ and then to build the scFv secondary structure. The scFv secondary structure was built using the Build Model command in Builder/Insight II. This command prepares Modeller input files to connect the VH and VL models by the linker $(Gly_4\text{-}Ser)_3$. Certain other commands were also used to build the linker $(Gly4\text{-}Ser)_3$, such as the Get command, which reads files containing single-letter amino acid codes; the Put command, which writes output to files of either single-sequence rows or full alignments; and the Copy command, which copies the amino acid sequence row. The last command used was the Start command, which starts the Modeller background job.

2.3. Energy Minimisation of scFv Predicted Structures

Insight II contains all of the necessary information to define the topology, coordinates, and force field parameters. These parameters include the atom types and partial charges. When doing energy minimization, the Discover module of Insight II provides a convenient interface. This module builds Discover input files from information provided through graphical interfaces, and it allows Discover jobs to run interactively.

In Insight II, the force field parameters were set up using three command steps: first, the Forcefield/Select

command was entered, and then atom types were assigned using the Fix command for Potential Action in Forcefield/Potentials. Alternatively, the atoms types were assigned with the Atom/Potential command in the Biopolymer module, and then the Accept option for Potential Action in Forcefield/Potentials was used. Finally, to assign the charges, the Fix command was used for both Partial Chg Action and Formal Chg Action under Forcefield/Potentials.

The next step was used to minimise the energy of the scFv antibody structure. The correctness of the structure has already been checked using the assigned atom types and partial charges commands. To perform this step, the command Potential or Partial charge in Molecule/Label was used to label each atom. The structural information was specified by moving to the Discover module in Insight II. The Constraint Pull-down menu contains various atom-constraining and restraining procedures. In Parameters, the simulation type for Discover (Minimize, Dynamics, etc.) and the choices for the cut-off parameters for non-bonded interactions were selected. Additionally, to start a simulation, the command Run/Run was entered for the object being calculated. Each Discover run was assigned a number based on the order of the execution start times. The files created during the execution were identified by the calculation object and the job integer, and the file extension specifies the file type.

2.4. Structural Evaluation of the Heavy-Chain Model and Light-Chain Model

To evaluate the scFv structures, Ramchandran plotting, Verify3D and ERRAT were used. These programmes are freely available at the UCLA-DOE server: (http://www.Shannon.mbi.ucla.edu/DOE/services/SV/). These structural evaluation methods allowed the reliable recognition of suitable templates for the heavy and light chains of the scFv protein structure. Additionally, the structural evaluation methods were able to produce sequence-structure alignments with fewer gaps.

Root mean square deviation (RMSD) is a technique that was developed by Giannakakos (2000). This method was used to evaluate the similarity of protein structures to their templates and to determine the accuracy of the alignment of the residues of two structures. The units used are Angstroms (Å).

$$MSD = \sqrt{\sum_{I=1}^{N} D_i^2 / N}$$

where,

i is the index that identifies a pair of corresponding residues in two structures.

N is the number of atoms. D_i is the distance between corresponding i atoms.

The computation of the RMSD requires a sequence alignment that defines which pairs of residues correspond to each other and an optimal superposition of the two structures in space.

3. RESULTS

3.1. VH and VL Chains Nucleotide and Amino Acids Sequences

The nucleotide and amino acid sequences of the VH and VL chains are shown in **Figures 1** and **2**. The DNA sequences of both the VH and VL chains were obtained from First BASE Laboratories Sdn. Bhd and translated into amino acid sequences in the TRANSLATE programme. The three CDRs for both chains were highlighted using KABAT numbering. The nucleotide and amino acid sequences of the VH and VL chains were obtained for use in model prediction. The CDR sequences are shown in red lettering in **Figure 2**.

3.2. Template Search and Selection

Generally, all current comparative modelling consists of four sequential steps: fold assignment and template selection, template-target alignment, model building and model evaluation. The selection of the template structure is generally performed by a programme that detects sequence similarity only, such as FASTA, BLAST, and programmes based on dynamic programming methods [8,9]. However, a slightly related sequence-structure pair needs to be identified through a more difficult method that relies on structural information or multiple sequences from the family of interest. First, a database search through unrelated sequence similarity searches was conducted by BSI-BLAST at the NCBI database http://www.ncbi.nlm.nih.gov.BLAST to identify a homologous protein that possessed a crystal structure for use as a template.

The X-BLAST identified many templates that were chosen to align with the VH and VL target chains, as shown in **Tables 1** and **2**. A reliable structure can only be obtained when the target and template are properly aligned. That state can only be achieved when the sequence identity between the modelled sequence and at least one known structure is >30% [10]. The heavy chain (VH) consisted of approximately 113 amino acids with 75% identity with 3BKY, the template sequence shown in **Table 1**. The amino acid sequence of the light chain (VL) consisted of approximately 105 amino acids with 85% identity with 1AY1, the template sequence shown in **Table 2**. The CDR regions in the VH and VL amino acids were determined using KABAT (www.kabatdatabase.com), as shown in **Figures 1** and **2**. The CDRs of the heavy chain are in boldface, with CDR-H1 shown in red, CDR-H2 in blue and CDR-H3 in


```
>gi|16124255|gb|AAG28706.2|  single chain antibody against rice stripe virus protein P20 [synthetic
construct]
Length=261

 Score =  245 bits (625),  Expect(2) = 3e-117
 Identities = 118/119 (99%), Positives = 118/119 (99%), Gaps = 0/119 (0%)
 Frame = +1

Query  1    MAQVKLQQ*GTEVVKPGASVKLSCKASGYIFTSYDIDWVRQTPEQGLEWIGWIFPGEGST  180
            MAQVKLQQ GTEVVKPGASVKLSCKASGYIFTSYDIDWVRQTPEQGLEWIGWIFPGEGST
Sbjct  1    MAQVKLQQXGTEVVKPGASVKLSCKASGYIFTSYDIDWVRQTPEQGLEWIGWIFPGEGST  60

Query  181  EYNEKFKGRATLSVDKSSSTAYMELTRLTSEDSAVYFCARGDYYRRYFDLWGQGTTVTV  357
            EYNEKFKGRATLSVDKSSSTAYMELTRLTSEDSAVYFCARGDYYRRYFDLWGQGTTVTV
Sbjct  61   EYNEKFKGRATLSVDKSSSTAYMELTRLTSEDSAVYFCARGDYYRRYFDLWGQGTTVTV  119


 Score =  201 bits (511),  Expect(2) = 3e-117
 Identities = 108/108 (100%), Positives = 108/108 (100%), Gaps = 0/108 (0%)
 Frame = +2

Query  410  SDIELTQSPAIMSASPGERVTMTCSASSSIRYIYWYQQKPGSSPRLLIYDTSNVAPGVPF  589
            SDIELTQSPAIMSASPGERVTMTCSASSSIRYIYWYQQKPGSSPRLLIYDTSNVAPGVPF
Sbjct  136  SDIELTQSPAIMSASPGERVTMTCSASSSIRYIYWYQQKPGSSPRLLIYDTSNVAPGVPF  195

Query  590  RFsgsgsgtsysLTINRMEAEDAATYYCQEWSGYPYTFGGGTKLELKR  733
            RFSGSGSGTSYSLTINRMEAEDAATYYCQEWSGYPYTFGGGTKLELKR
Sbjct  196  RFSGSGSGTSYSLTINRMEAEDAATYYCQEWSGYPYTFGGGTKLELKR  243
```

Figure 1. DNA Blast in NCBI GenBank of the investigated scFv gene. The identity was 99% for the heavy chain of the scFv gene and was 100% for the light chain of the scFv gene gi[1612455[gb[AAG28706.2].

yellow. The CDRs of the light chain are also in boldface, with CDR-L1 shown in red, CDR-L2 in yellow and CDR3-L3 in green. The similarity between two corresponding amino acids in the sequence alignment of the target chains and their templates in this work was very high; therefore, the predicted structures were accurate and reliable.

3.3. Target-Template Alignments

Multiple sequence alignment is useful for placing deletions or insertions in areas where the sequences are significantly different [11]. The structural information from the template structure can also be used to guide the alignment by modifying the gap penalty function to favour gaps in structurally reasonable contexts. The VL and VH chain models were further aligned with the template sequences by box-shading the conserved regions to elucidate the variability of the amino acids that conferred certain differences between the sequences. The target domains that were assessed to interact through the interface modes in a given PDB structure were listed as candidate members of the heavy- and light-chain complex, as shown in **Tables 1** and **2**. **Figure 3(a)** shows several amino acid variations and insertion regions, especially between the heavy-chain amino acid sequence alignment with the 3BKY template and as shows in **Figure 3(b)** the light-chain amino acid alignment with the 1AY1 template.

Comparative protein modelling stresses that accuracy can be higher if the segments of the model are selected from homologous sequences (Blundell and Srinivasan 1996). High identity between the target and template sequences generally allows the construction of a predicted 3D structure with high accuracy. An identity of above 60% tends to produce a structure comparable to medium-resolution NMR or low-resolution crystallography without crystallisation or experimental structure determination [12]. Because homology modelling was used to produce the structural models in this work, no crystallisation or experimental structural determination was needed. Additionally, the numbers of structurally conserved regions (SCRs), comprising approximately 85% of the light chain and 75% of the heavy chain, were identified, and the accuracy of the predicted structure was high and reliable. The 3D structures predicted for the light chain and heavy chain were constructed through the SWISS-MODEL website (http://swissmodel.expasy.org/workspace/index), as shown in **Figures 4(a)** and **(b)**.

3.4. Building the Full Structure of the scFv Antibody Using Builder/Insight II Software

Builder/Insight II software was used to connect the VH and VL models by the linker $(Gly4\text{-}Ser)_3$ and then to build the full scFv secondary structure in CPK display, as shown in **Figures 5(a)-(c)**. The CPK model shows all of the CDRs on the surface of the molecule. The peptide linker appears in the middle of the structure, whereas

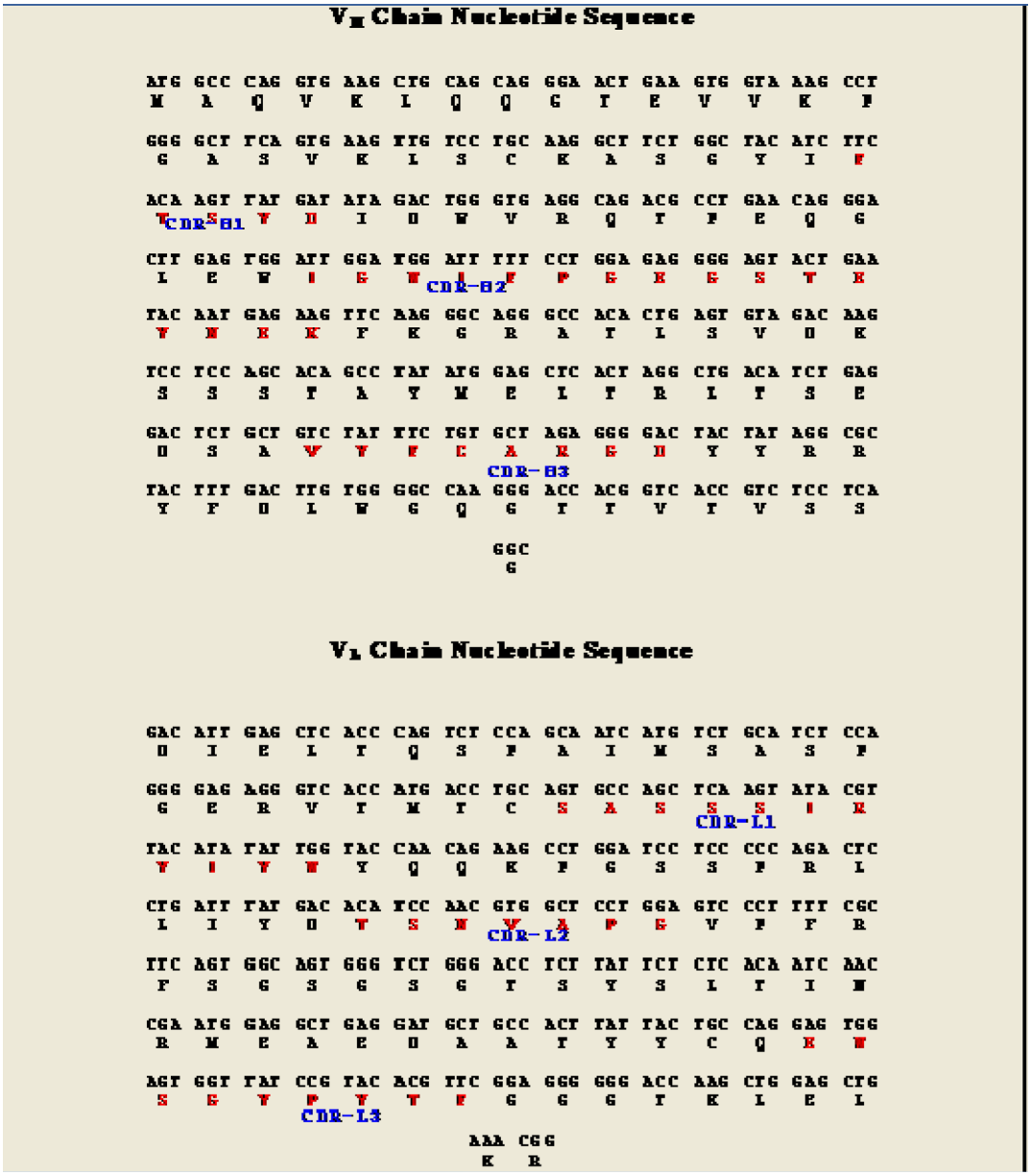


Figure 2. The nucleotide and amino acid sequences for the VH and VL chains were determined to use later in model prediction. The sequences were obtained from First BASE Laboratories Sdn. Bhd and translated using the TRANSLATE programme.

Table 1. Selecting target templates for the heavy chain.

pdb|3DIF|B	Crystal Structure of Fabox117	75%
pdb|3BKY |H	Crystal Structure of the heavy chain of Chimeric Antibody C2 (**the chosen template**)	75%
pdb|3HQK|Q	Q Chain Q, X-Ray Crystal Structure of An Arginine Ag	73%
pdb|2OSL|H	Chain H, Crystal Structure of Rituximab Fab	76%

CDR-HI, CDR-H2 and CDR-H3 are found in the upper part of the structure, and CDR-L1, CDR-L2, and CDR-L3 are at the bottom of the structure. The linker provides the molecular flexibility required to move 35 to 40 Å (10^{-9} Kcal/mol) [1]. The root mean square deviation (RMSD) evaluation method was used to measure the accuracy of the loop structures, such as the linker $(Gly_4\text{-}Ser)_3$ and the sequences in the VH and VL insertion gaps. The insertion of gaps into an alignment between two protein sequences, known as the loop struc-

Table 2. Selecting target templates for the light chain.

pdb|1AY1|L	Chain L, Anti Taq Fab Tp7 (**the chosen template**)	85%
pdb|1BGX|L	Chain L, Taq Polymerase In Complex With ligand bound Tp7, An Inhibitory Fab	85%
pdb|1BAF|L Chain	L, 2.9 Angstroms Resolution Structure Of An Anti-Dinitrophenyl-Spin-Label Monoclonal Antibody Fab Fragment With ligand bound	86%

```
Query  7    QVKLQQ*GTEVVKPGASVKLSCKASGYIFTSYDIDWVRQTPEQGLEWIGWIFPGEGSTEY  186
            +VKLQQ G E+VKPGASVK+SCKASGY FTSY I WV+Q P QGLEWIGW+FPG G+T+Y
Sbjct  1    EVKLQQSGPELVKPGASVKISCKASGYSFTSYYIHWVKQRPGQGLEWIGWVFPGSGNTKY  60

Query  187  NEKFKGRATLSVDKSSSTAYMELTRLTSEDSAVYFCARGDYYRRYFDLWGQGTTVTV  357
            NEKFKG+ATL+ D SSSTAYM+L+ LTSEDSAVYFCARG+Y R +F  WGQGT VTV
Sbjct  61   NEKFKGKATLTADTSSSTAYMQLSSLTSEDSAVYFCARGNYDRAWFAYWGQGTLVTV  117
```

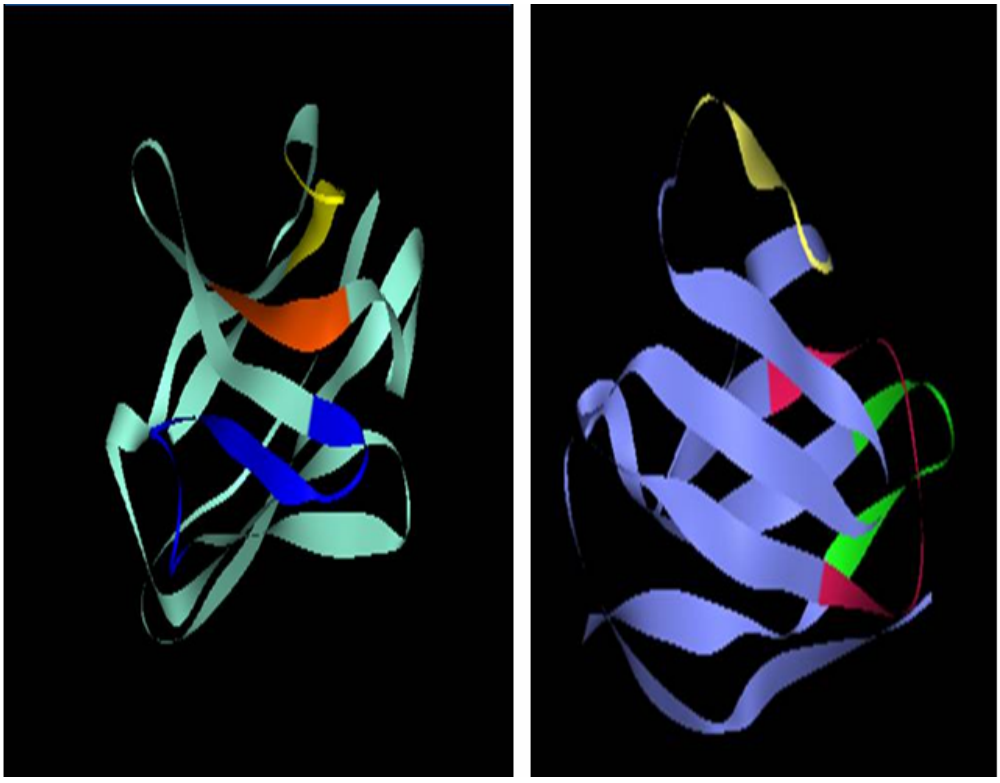
(a)

```
Query  5    DIELTQSPAIMSASPGERVTMTCSASSSIRYIYWYQQKPGSSPRLLIYDTSNVAPGVPFR  184
            DI++TQSPAIMSASPGE+VTMTCSASSS+ Y+YWYQQKPGSSPRLLIYD++N+A GVP R
Sbjct  1    DIQMTQSPAIMSASPGEKVTMTCSASSSVSYMYWYQQKPGSSPRLLIYDSTNLASGVPVR  60

Query  185  FsgsgsgtsysLTINRMEAEDAATYYCQEWSGYPYTFGGGTKLELKR  325
            FSGSGSGTSYSLTI+RMEAEDAATYYCQ+WS YP TFG GTKLELKR
Sbjct  61   FSGSGSGTSYSLTISRMEAEDAATYYCQQWSTYPLTFGAGTKLELKR  107
```

(b)

Figure 3. (a) Heavy-chain sequence alignment with the 3BKY template in the ncbi-blast website. The sequence identity was 75%. This result was then used for the heavy-chain model prediction. (b) Light-chain sequence alignment with the 1AY1 template sequence in the ncbi-blast website. The sequence identity was 85%. This result was then used for the light-chain model prediction.



(a) (b)

Figure 4. The heavy- and light-chain 3D structures for the scFv antibody and its CDRs. (a) Heavy chain: CDR-HI (red), CDR-H2 (blue) and CDR-H3 (yellow). (b) Light chain: CDR-Ll (red), CDR-L2 (yellow) and CDR-L3 (green). The CDR amino acid regions in the heavy chain (VH) and light chain (VL) were determined using KABAT numbering.

ture, is a major determinant of the accuracy of the alignment.

3.5. Energy Minimisation of the Predicted Structures

In the energy minimisation of the protein, the hydrogen atoms were relaxed first, followed by the side chains of the amino acid residues, and finally the whole molecule. Despite the logic of this approach, however, the structures minimised by an unconstrained path fit the experimental structures better than those minimised by constrained paths. Moreover, the unconstrained path

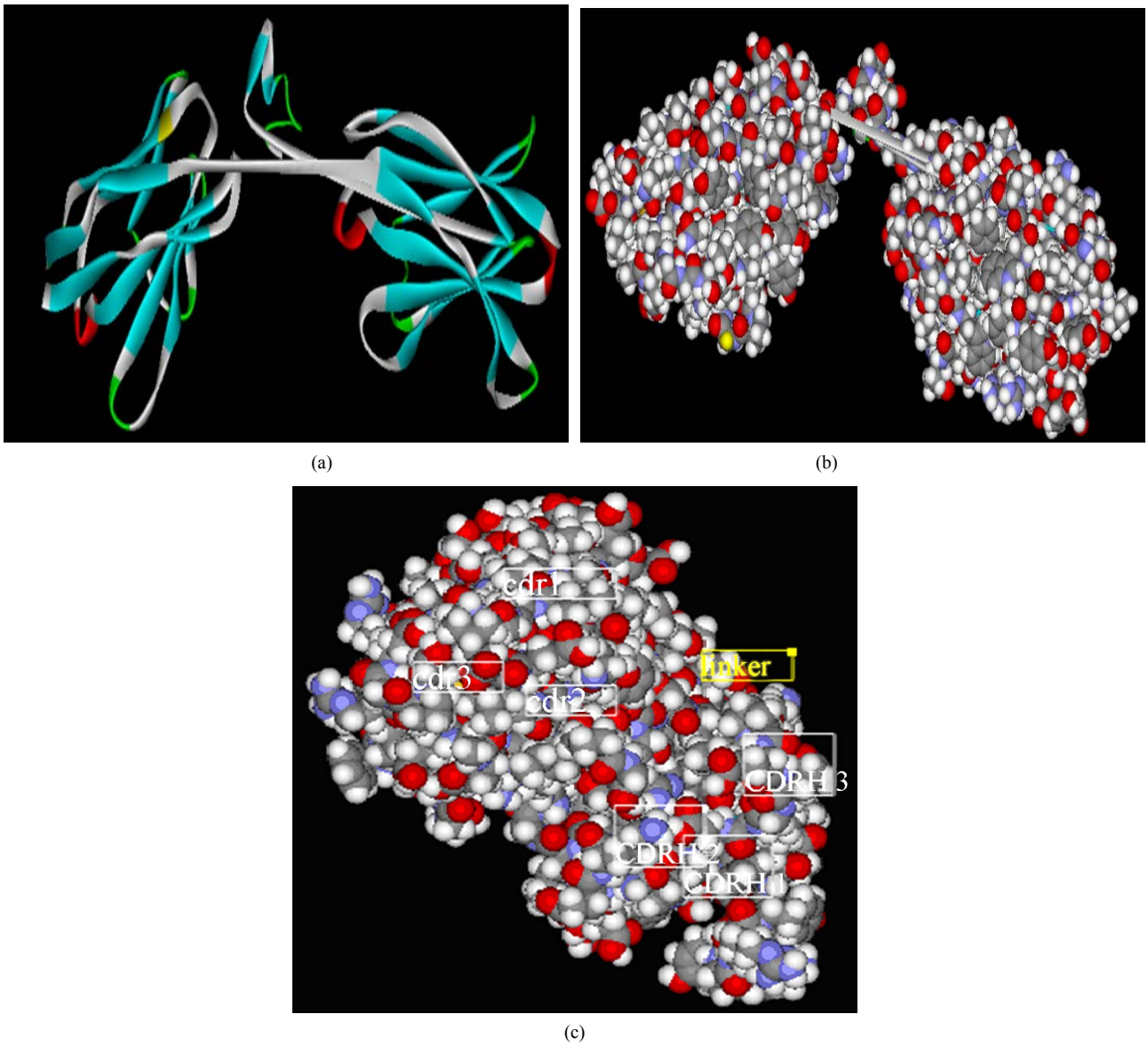


Figure 5. (a) The full scFv protein model built by joining the VH and VL chains together by the peptide linker $(Gly_4\text{-}Ser)_3$ using BUILDER/Insight II. (b) The full scFv protein model is shown in CPK display. The heavy-chain, linker and light-chain models are clearly shown. (c) The full scFv protein model is shown in CPK display. The CPK model was energy-minimised in a CFF91 force field. The CDR molecules are shown on the surface of the CPK model.

required much less computer time. The effects of the steepest descents were compared with those of the conjugate gradient algorithms in energy minimisation. Finally, steepest descents were used in the initial stages of the minimisation and conjugate gradients in the final stages of the minimisation. The full scFv model was energy-minimised using 30 steps of steepest descent followed by 50 steps of conjugate gradient in the water shell, calculated with Amber 6.0 (University of California, USA) with certain restraints to preferred geometric regions; also, two Na^+ ions were added to neutralise the system.

3.6. Structural Evaluation of the Heavy-Chain and Light-Chain Models

A knowledge-based homology modelling approach was used to predict the 3D structures of the heavy and light chains. The templates of the predicted structures were evaluated using three independent evaluation methods to gain confidence about the correctness and accuracy of the templates. All of the templates were submitted to the structure evaluation website (UCLA-DOE). The structures were evaluated using three programmes, Ramchandran plotting [6], Verify3D [7] and ERRAT. These methods were essential for understanding 3D protein

models and the estimation of their accuracy. Both the overall accuracy and the accuracy in the individual regions of a model must be determined.

The predicted structures of the VH and VL chains met the above standard, as x-blast expanded the set of homologues of the target sequence, and the scoring matrix was used to search for new homologues. Additionally, template sequences with high identities to the target sequences were used, specifically 99% identity for the heavy chain and 100% for the light chain. The high sequence identity ensured a high accuracy for the models because the average structural similarity increases with sequence identity.

3.6.1. ERRAT Method

As shown in **Figure 6(a)**, ERRAT is a programme for verifying protein structures that have been determined by crystallography [13]. It is also useful for verifying protein structures from the numbers of non-bounded contacts within a cut-off distance of 3.5 Å between different pairs of atom types (CC, CN, CO, NN, NO, OO). The error function is based on the statistics of non-bound atom-atom interactions in the reported structure compared with high-resolution structures. As shown in **Figure 6(a)**, the predicted structure of the light chain exhibited an overall quality factor of 78.505%. Additionally, in **Figure 6(b)**, two lines were drawn to indicate the confidence with, which it was possible to reject regions that exceed the error value. The predicted models show- ing high resolution in the crystal structure generally produce values of approximately 70% [14]. The confidence level of an overall quality factor for the heavy chain of 70.347% significantly determined the correctness of the predicted structure (**Figure 6(b)**). The model evaluation method outperforms the programmes in the high sequence identity range, producing good modelling accuracy overall.

3.6.2. Verify3D Method

Verify3D evaluates the environment of each residue in a model with respect to the expected environment, as found in high-resolution X-ray structures [15]. Verify3D analyses the compatibility of an atomic model (3D) with its own amino acid sequence (1D) [16]. The accuracy of a 3D model can be assessed by its 3D profile, regardless of whether the model has been produced by X-ray, NMR or computational procedures, by comparing the model to its amino acid sequence using its 3D profile [17]. The 3D-1D average score against sequence number, as indicated in **Figure 7(a)**, shows that 100% of the total residues scored from 0.2 to 0.7 in the light chain, whereas 88.5% of the total residues scored from 0.15 to 0.7 in the heavy chain. As shown in **Figure 7(b)**, both predicted models have 3D-1D average scores of more than 0.15. These models contain high-scoring regions, with the correctness of the good models above 0.15. The results significantly determined the correctness of the model as the average score of distinct structures. The average is often a score below 0.1 that may dip below zero at its lowest points [17].

3.6.3. Ramchandran Plot Method

The Ramchandran plot method tests the light- and heavy-chain polypeptide angles and identifies favoured residues and allowed residues. In the light-chain predicted model, the test showed that 83.0% (73) of the

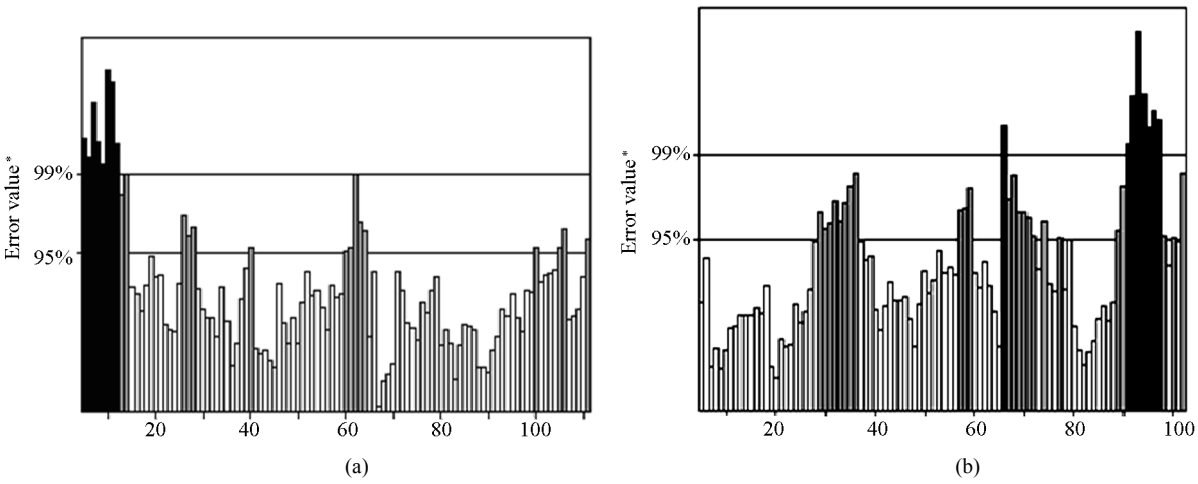
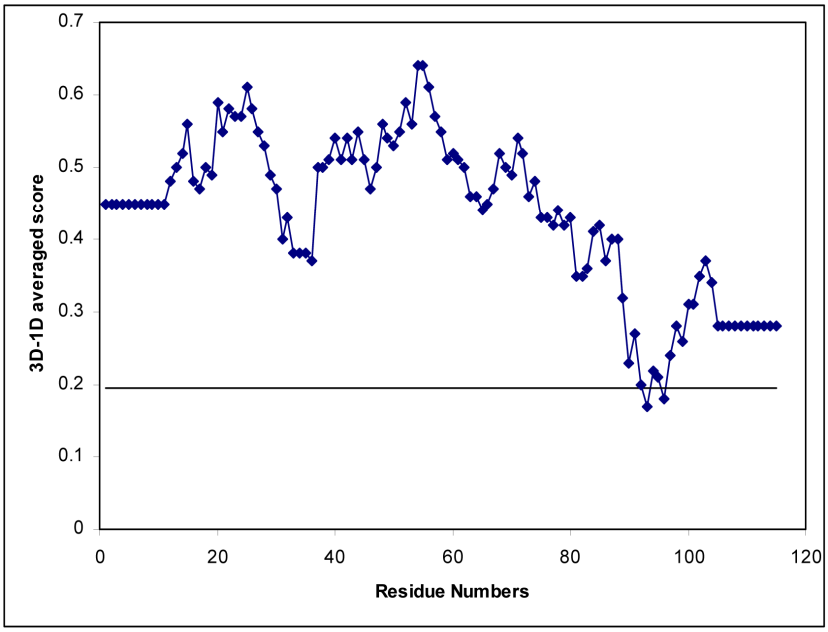
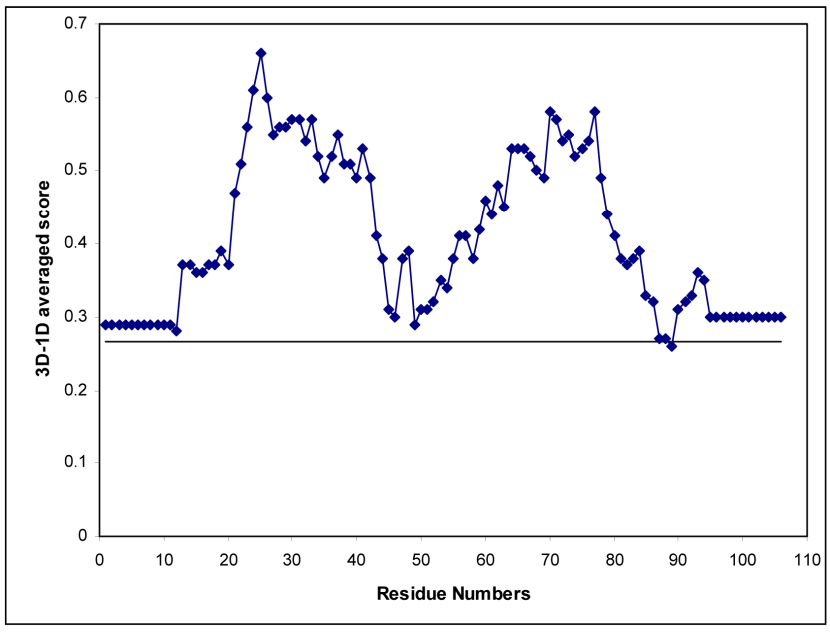


Figure 6. (a) The ERRAT evaluation methods for the light-chain residues gave 78.505% as the overall quality factor; this ERRAT value is considered good enough to use this model. In the ERRAT histogram, the correct regions are shown in black, and the incorrect regions are shown in grey. (b) The ERRAT evaluation method for the heavy-chain residues gave 70.347% as the overall quality factor; this ERRAT value is considered good enough to use this model. In the ERRAT histogram, the correct regions are shown in black, and the incorrect regions are shown in grey.



(a)



(b)

Figure 7. (a) The Verify3D curve for the light-chain model execs between residue numbers and 3-1 dimensions score. The light-chain model gave more than 86%. The residues of the light-chain model scored 0.3 of 3D-ID. (b) The Verify3D curve of the heavy-chain model execs between residue numbers and 3-1 dimensions score. The heavy-chain model gave more than 85%. The residues of the heavy-chain model scored more than 0.3 of 3D-ID.

residues lie in the most favoured region, with 14.8% (13) of the residues in the additional allowed region, as shown in **Figure 8(a)**. The quality of the plot was better than that of the template 1AY1, as only 78.0% and 21.0% of the residues of the template structure 1AY1 fell into the most favoured region and additional allowed region, respectively. However, 2.3 residues were in the disallowed region for both models. The catalytic serine residue (Ser, Gly and Met as 113, 114, and 116, respectively) lies in the most favoured region. This standard, described by [18], was a typical conformation for the nucleophilic elbow, which was located in the tightly constrained

beta-turn-type structure between a beta-strand and an alpha-helix. The Ramachandran plot of the heavy-chain predicted model, as shown in **Figure 8(b),** reveals that 81.8% (81) of the residues lie in the most favoured region, with 16.2% (16) of the residues in the additional allowed region. The quality of the plot was better than that of the template 3BKY, as only 76.0% and 23.0% of the residues of the template structure 3BKY fell into the most favoured region and additional allowed region, respectively. However, zero residues were in the disallowed region for both models. The catalytic serine residue (Ser113) lies in the allowed region.

4. DISCUSSION

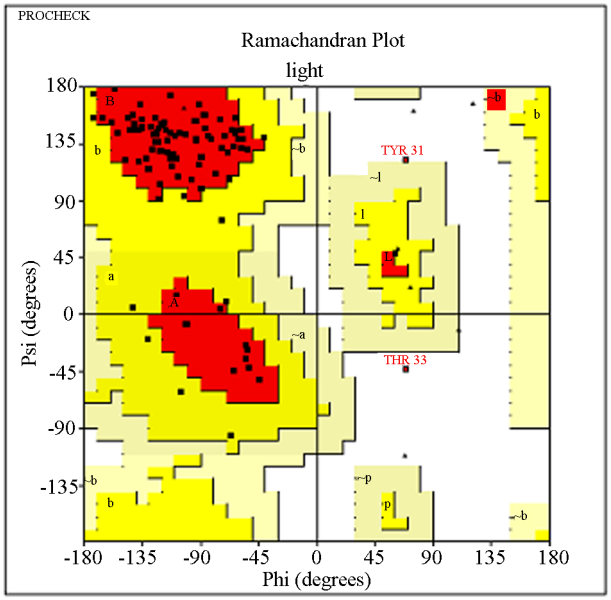
Knowledge-based homology modelling relies on the identification of one known protein structure, which is likely to resemble the structure of the query sequence, and on the production of an alignment that maps the residues in the query sequence to the residues in the template sequence. Therefore, the heavy- and light-chain genes were sequenced, and the sequences were deposited in GenBank. The mapped residues in the query were aligned to residues in the template sequence. A number of scFv structures at the Protein Data Bank (PDB) www.rcsb.org/pdb were used [19], and general information on antigen binding was documented. Hence, the scFv protein sequences in PDB were used in x-BLAST to identify suitable templates for homology modelling. **Figure 3(a)** shows the amino acid sequence alignment of the light-chain predicted structure and the 1AY1 template. Additionally, **Figure 3(b)** shows the amino acid sequence alignment of the heavy-chain predicted structure and the 3BKY template structure. There were a few amino acid variations, and there were several insertion regions, especially between the heavy-chain predicted structure and the 3BKY template structure, as shown in **Figure 3(b)**.

Normally, an optimal alignment leads to a more accurate model. The PDB search results showed a high sequence similarity of 75% for 3BKY, the heavy-chain template, and of 85% for 1AY1, the light-chain template. Any model can be predicted with sequence similarity equal to or greater than 30% [10]. Thus, the availability of a structural homolog at PDB was confirmed. The scFv antibody sequence was then submitted to SWISS-MODEL, and the VH and VL structures were separately modelled. As shown in **Figures 4(a)** and **(b)**, mapping the complimentary-determining regions (CDRs) are important for supporting library diversity [20]. The canonical conformations for the CDRs in the scFv antibody 3D structure were successfully mapped. The CDRs, as shown in **Figures 4(a)** and **(b)**, were mapped to identify their positions in the heavy and light chains. The VH and VL models were linked by the synthetic peptide $[(GlY_4Ser)_3$, followed by energy minimisation in a CFF91 force field. The modelled scFv structure was represented as a CPK model, and the CDRs were mapped with Accelrys Visualize at the website (http://www.accelrys.com). Thus, the CDRs in the modelled antibody structure were determined by KABAT numbering (**Figures 4(a)** and **(b)**).

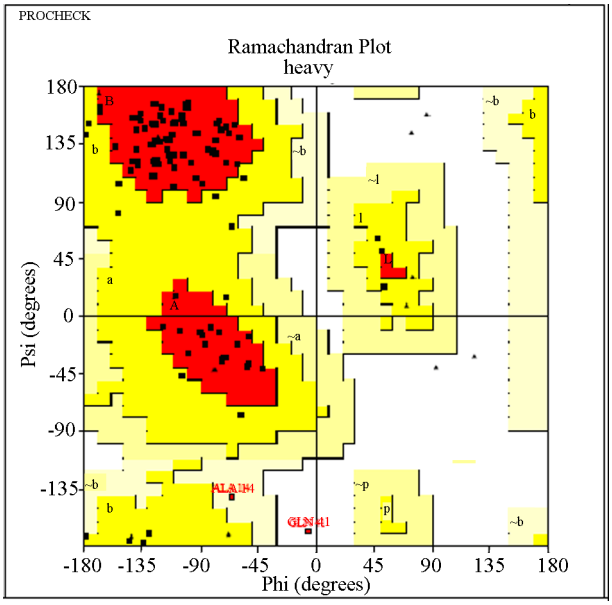
The loop region of the structure was the most important task in modelling the scFv protein. The loop regions of the model are the structures constructed without a template guide [21,22]. The loop evaluation was measured using the root mean square deviation (RMSD). Therefore, the synthetic peptide $(GlY_4Ser)_3$ that was built using BUILDER/Insight II had to be measured. Also, the loop structure was recorded with the root mean square deviation (RMSD), which was 4 Kcal/mol in the heavy chain and 2 Kcal/mol in the light chain. An optimal superposition (minimal RMSD) can be achieved by translating and rotating one structure to its relative structure in space [23]. Therefore, the optimal alignment (optimal set of pairs of corresponding residues) was obtained and is given in **Tables 1** and **2**. As expected for structures of good quality, the templates of the correct models have average energy profiles smaller than zero over most of their lengths. The models based on incorrect alignments show higher energy compared with reliable structures.

These results confirm the efficiency of the achieved minimisation strategy in modelling closely related homologies. To determine the reliability of the united atom approximation, all of the above minimisations were performed with united atom models. This approximation gave structures with similar but slightly higher RMS deviations than the all-atom models, but gave additional savings of 60% - 70% in computer time. Previously, steepest descents have been used in the initial stages of minimisation and conjugate gradients in the final stages of minimisation. Therefore, the structures minimised by conjugate gradients alone resembled the structures minimised initially by the steepest descents and subsequently by the conjugate gradient algorithms.

The predicted VL and VH structures were evaluated using three independent evaluation methods to gain confidence about the correctness of the predicted structures. Also, the evaluation of a model normally involves checking the sequence identity and functional environment [15]. The VH and VL structures were evaluated using Ramachandran plots, Verify3D and ERRAT. These methods are freely available at the UCLA–DOE server (www.Shannon.mbi.ucla.edu). Furthermore, Hatem *et al.* [14] reported that very good models score above 70% with ERRAT evaluation methods; thus, in this work, the correctness of both predicted structures was significantly above this confidence level, with scores of 78.505% for the light chain and 70.347% for the heavy chain. Moreover, in Verify3D, in which the method analysed the compatibility of an atomic model



(a)



(b)

Figure 8. (a) A Ramachandran plot showing the analysis of 118 structures with a resolution of at least 2.0 Angstroms and an R-factor no greater than 20%. A good quality model would be expected to have over 90% of the residues in the most favoured regions. In this model, more than 90% of the residues are in the favoured regions. (b) A Ramachandran plot showing the analysis of 118 structures with a resolution of at least 2.0 Angstroms and an R-factor no greater than 20%. A good quality model would be expected to have over 90% of the residues in the most favoured regions. In the heavy-chain model, more than 99% of the residues are in the favoured region.

(3D) with its own amino acid sequence (1D) [16], the light-chain and heavy-chain residues scored more than 0.3 of 3D-ID in that method, as shown in **Figures 7(a)** and **(b)**. Therefore, the results determined that both models were correct models that could be predicted with the templates used, as approved by Hatem *et al.* [14].

A basic requirement for a good model is the stereochemistry in displaying main-chain torsion angles phi, psi (φ, ψ) as determined by procheck [21]. procheck is widely used to calculate the Ramchandran angles of protein structures, particularly crystal structures available in the Protein Data Base (PDB) [9]. In the Ramchandran method, the polypeptide chain is displayed using the φ, ψ angles pair in a given protein structure as described by Ramchandran [6]. In this paper, the models were considered to be good quality because 99% and 90% of the heavy- and light-chain residues were in the favoured regions, as shown in **Figures 8(a)** and **(b)**, respectively. Moreover, none of the residues were in the disallowed region for either model. The Ramchandran plot is less effective than Verify3D at revealing damaged fragments, as it sometimes appears normal even though the structure is completely wrong.

5. CONCLUSION

The study of the scFv protein prediction models was accurate enough to be useful in essential ligand characterisation. This work presents the anti-MCF-7 scFv protein sequence against PDB (protein database), using BLAST-P to identify suitable templates for homology modelling. The PDB search results show a high sequence similarity (99%) to a synthetic peptide. Thus, the availability of a structural homolog at PDB was confirmed. Next, the anti-MCF-7 scFv amino acid sequence was submitted to SWISS-MODEL, and the VH and VL structures were separately modelled. The models were represented as ribbons, generated using RasMol. The canonical conformations for the CDRs in scFv anti-MCF-7 are mapped in 3D and mapped regions. The individually modelled VH and VL structures were linked by a synthetic peptide [$(Gly4Ser)_3$ using BUILDER/Insight II, followed by energy minimisation in a CFF91 force field. The modelled anti-MCF-7 scFv structure is represented as a CPK model, and the CDRs are mapped. Thus, the structure of an anti-MCF-7 scFv was modelled, and the CDRs were mapped to the structure in 3D. The model was subsequently evaluated using Verify3D, ERRAT and Ramachandran plots. Parts of the protein with unsatisfactory energy were realigned to the template, and the whole process of model building and evaluation was repeated until most of the average energy profile was below zero.

REFERENCES

[1] Chua, K.H., *et al.* (2006) Bioinformatics in molecular

immunology laboratories demonstrated: Modeling an anti-CMV scFv antibody. *Bioinformation*, **1**, 118-120.

[2] Blundell, T.L. and Srinivasan, N. (1996) Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proceedings of National Academy of Sciences of the USA*, **93**, 14243-14248. doi:10.1073/pnas.93.25.14243

[3] Katchalski, K., Katzir, E., Shariv. I., *et al.* (1992) Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of National Academy of Sciences of the USA*, **89**, 2195-2199.

[4] Wu, S. and Zhang, Y. (2009) Chapter 11: Protein structure prediction. In: D. Edwards, *et al.*, Eds., *Bioinformatics*: *Tools and Applications*, Springer Science+Business Media, LLC, Berlin, 225-242.

[5] Janin, J., Henrick, K., Moult, J., *et al*. (2003) A critical assessment of predicted interactions. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **52**, 2-9. doi:10.1002/prot.10381

[6] Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, **7**, 95-99. doi:10.1016/S0022-2836(63)80023-6

[7] Eisenberg, D., Lüthy, R. and Bowie, J.U. (1997) VERIFY3D: Assessment of protein models with three-dimensional profile. *Methods in Enzymology*, **277**, 396-404. doi:10.1016/S0076-6879(97)77022-8

[8] Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proceedings of National Academy of Sciences of the USA*, **95**, 13597-13602. doi:10.1073/pnas.95.23.13597

[9] Dlakic, M. (2002) A model of the replication fork blocking protein Fob1p based on the catalytic core domain of retroviral integrates protein. *Protein Science*, **11**, 1274-1277.

[10] Marti-Renom, M.A., Stuart, A.C., Fiser, A., *et al.* (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, **29**, 291-325. doi:10.1146/annurev.biophys.29.1.291

[11] Madhusudhan, M.S., *et al.* (2006) Variable gap penalty for protein sequence-structure alignment. *Protein Engineering*, *Design and Selection*, **19**, 129-133. doi:10.1093/protein/gzj005

[12] Sanchez, R., Pieper, U., Melo, F., *et al.* (2000) Protein structure modeling for structural genomics. *Nature Structural Biology*, **7**, 986-990.

[13] Colovos, C. and Yeates, T.O. (1993) Verification of protein structures: Patterns of non-bonded atomic interactions. *Protein Science*, **2**, 1511-1519. doi:10.1002/pro.5560020916

[14] Hatem, R., Pierre, B., Elie, E., *et al*. (2005) Structural and functional analysis of the C-terminal STAS (sulfate transporter and anti-sigma antagonist) domain of the *Arabidopsis thaliana* sulfate transporter SULTR. *The Journal of Biological Chemistry*, **280**, 15976-15983.

[15] Fiser, A., Sanchez, R., Melo, F., *et al*. (2001) Comparative protein structure modeling. In: Becker, O.M., Ed., *Computational Biochemistry and Biophysics*, Marcel Dekker, New York, 275-312.

[16] Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170. doi:10.1126/science.1853201

[17] Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-85. doi:10.1038/356083a0

[18] Tyndall, D.A., Linda, A., Fothergill-Gilmore, P., *et al*. (2000) Crystal structure of a thermostable lipase from Bacillus stearohermophilus P1. *Journal of Molecular Biology*, **323**, 859-869.

[19] Bernstein, F.C., Koetzle, T.F., Williams, J.B., *et al*. (1977) Protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**, 535-542. doi:10.1016/S0022-2836(77)80200-3

[20] DeBartolo, J., Colubri, A., Jha, A.K., *et al*. (2009) Mimicking the folding pathway to improve homology free protein structure prediction. *Proceedings of National Academy of Sciences of the USA*, **106**, 3734-3739.

[21] Laskowski, R.A., Moss, D.S. and Thornton, J.M. (1993) Main-chain bond lengths and bond angles in protein structures. *Journal of Molecular Biology*, **231**, 1049-1067.

[22] Vriend, G. (1990) WHATIF: A molecular modeling and drug design program. *Journal of Molecular Graphics*, **8**, 52-56. doi:10.1016/0263-7855(90)80070-V

[23] Ryan, D., Xiaotao, Q., Rosemarie, S., *et al*. (2011) Relative packing groups in template-based structure prediction: Cooperative effects of true positive constraints. *Journal of Computational Biology*, **18**, 17-26.

LIST OF ABBREVIATIONS

scFv—Single Chain Fragment Variable
VL—Hyper variable light chain
VH—Hyper variable heavy chain
MCF—7-mammary gland carcinoma of the breast cells line
anti-CMV—anti-cucumber mosaic virus
CDRs—complementary determining regions
Gly—Glycine amino acid
Ser—Serine amino acid.

