Research in Pharmaceutical Sciences, February 2023; 18(1): 67-77
Received: 11-02-2022
Peer Reviewed: 28-02-2022
Revised: 06-03-2022
Accepted: 27-11-2022
Published: 24-12-2022

School of Pharmacy & Pharmaceutical Sciences
Isfahan University of Medical Sciences

Original Article

Preparing a database of corrected protein structures important in cell signaling pathways

Samaneh Hatami[1], Hajar Sirous[2], Karim Mahnam[3], Aylar Najafipour[1], and Afshin Fassihi[1,*]

[1]Department of Medicinal Chemistry, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, I.R. Iran.
[2]Bioinformatics Research Center, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, I.R. Iran.
[3]Department of Biology, Faculty of Science, Shahrekord University, Shahrekord, I.R. Iran.

Abstract

Background and purpose: Precise structures of macromolecules are important for structure-based drug design. Due to the limited resolution of some structures obtained from X-ray diffraction crystallography, differentiation between the NH and O atoms can be difficult. Sometimes a number of amino acids are missing from the protein structure. In this research, we intend to introduce a small database that we have prepared for providing the corrected 3D structure files of proteins frequently used in structure-based drug design protocols.
Experimental approach: 3454 soluble proteins belonging to the cancer signaling pathways were collected from the PDB database from which a dataset of 1001 was obtained. All were subjected to corrections in the protein preparation step. 896 protein structures out of 1001 were corrected successfully and the decision on the remained 105 proposed twelve for homology modeling to correct the missing residues. Three of them were subjected to molecular dynamics simulation for 30 ns.
Findings / Results: 896 corrected proteins were perfect and homology modeling on 12 proteins with missing residues in the backbone resulted in acceptable models according to Ramachandran, z-score, and DOPE energy plots. RMSD, RMSF, and Rg values verified the stability of the models after 30 ns molecular dynamics simulation.
Conclusion and implication: A collection of 1001 proteins were modified for some defects such as adjustment of the bond orders and formal charges, and addition of missing side chains of residues. Homology modeling corrected the amino missing backbone residues. This database will be completed for quite a lot of water-soluble proteins to be uploaded to the internet.

Keywords: PDB; Homology modeling; Molecular dynamics simulation; Protein database; Protein structure.

INTRODUCTION

Structure-based drug design protocols in modern medicinal chemistry such as molecular docking, molecular dynamics (MD) simulation, and structure-based virtual screening rely on the information obtained from the interactions of the ligand with its target macromolecule (1). These macromolecules are usually proteins, protein derivatives such as glycoproteins, and nucleic acids such as ribonucleic acid and deoxyribonucleic acids. To model these interactions, the three-dimensional (3D) structure of the macromolecule must be available. These structures give the researchers precise insight into the topology and geometry of the molecule. To design effective modulators of an enzyme or receptor, one needs to know the details of active/binding site features.

Access this article online	
	Website: http://rps.mui.ac.ir
	DOI: 10.4103/1735-5362.363597

*Corresponding author: A. Fassihi
Tel: +98-3137927100, Fax: +98-*3136680011*
Email: fassihi@pharm.mui.ac.ir

Over the past few decades, the 3D structures of more than a hundred thousand macromolecules have been determined using the X-ray diffraction of their crystal structures (2). Obtained data is overseen by an organization called the Worldwide Protein Data Bank (PDB) and is archived in freely available databases on the internet such as PDBe (3), PDBj (4), RCSB (5), and BMRB (1).

Precise and high-quality structures of these macromolecules along with accurate soft-wares are basic requirements for successful drug design procedures (6). There are always numerous defects in the 3D files acquired from databases. Some of these deficiencies seem to be small while some have great impacts on the spatial structure of the protein and some of its physicochemical features. As some examples, due to the limited resolution of some structures obtained from X-ray diffraction crystallography, differentiation between the NH and O atoms can be difficult and these faults must be corrected and replaced. There are numerous cases in that a number of amino acids are missed from the protein structure, sometimes a few and in a limited region, but occasionally in a considerable amount in different parts of the structure. The presence of these gaps along the amino acid sequence may cause lots of errors in structure-based drug design procedures. Docking simulations might be biased by the absence of some protein moiety. This can be attributed to different causes. It is probable that the missing amino acids belong to the active/binding site of the under-study drugs. Even if they are missed from regions away from the binding site, they can be still effective in their spatial structure. On the other hand, a misinterpretation of electrostatic potentials at the protein surface is inevitable when some amino acids are missing from the protein structure. Thus, it is not far from the expectation that the ligand-binding energies calculated for such proteins might be incorrect. Furthermore, statistical trends mined from the PDB, especially those concerning the protein surface, might be inaccurate if the absence of some residues is not properly handled.

While working with a crystal structure, researchers usually spend remarkable time and effort correcting the structural faults before starting molecular modeling calculations and it is usually a major problem for the end-users of the PDB data (7). The most important corrections are the addition of missing heavy atoms, hydrogen atoms, building missing loops, converting non-standard residues to their standard equivalents, correction of bond orders and formal charges for metal cofactors and adjacent atoms in metalloproteins, the addition of missing disulfide bonds between cysteine sulfhydryl groups, the addition of side chains or amino acids missing from the structure, ionization and protonation states of the histidine and also aspartic acid and glutamic acid units. Usually, after all these modifications and corrections the protein structures are optimized energetically (8-13).

The importance of using accurate protein initial structures for molecular modeling and structure-based virtual screening calculations has led some research groups to focus on providing appropriate databases in this area (10). To the best of our knowledge, there are just a few databases available on the internet regarding such a subject. In this regard, the database entitled mpstruc "membrane protein of the known 3D structure" (https://blanco.biomol.uci.edu/mpstruc/), developed by the research team of Professor Stephen H. White, at the University of California can be mentioned.

Cancer is one of the most common causes of death in human societies (14). There are many protein targets in the cancer signaling pathway that can be used in the design of antineoplastic drugs (15). PDB files of these proteins are available but similar to many such files; these have the above-mentioned defects.

Here, in this research, we intend to introduce a small database, which we have prepared for providing the corrected three-dimensional structure files of proteins frequently used in structure-based drug design protocols. We started this project by modification of some soluble proteins in the cancer signaling pathway, but this database will be completed for quite a lot of water-soluble proteins in predicted scheduled steps to be uploaded to the internet. Many of the defects mentioned in this introduction will be resolved using proper soft-wares and methods.

MATERIALS AND METHODS

Instrumentation

All computations in this study were performed using a molecular modeling package from Schrodinger's Drug Discovery Suite 201 (Schrodinger, Inc., LLC, New York, USA) installed on an Intel Xeon CPU E5-2620 v2 @ 3.30 GHz, 64 GB RAM with 12 processors and a 2 GB graphics card of NVIDIA Quadro K2200 running Ubuntu 10.04 LTS (long-term support) as the operating system. Access to the Schrodinger modules as well as the capability to organize and analyze data was provided by Maestro as a portal interface of Schrodinger.

Protein datasets

In the first step, some of the protein families that play the most important roles in cancer signaling pathways were searched on the Protein PDB website. Then, the pdb codes belonging to the protein structures resolved using the X-ray diffraction method were collected. For each protein family, an Excel spreadsheet comprising some key information and features was prepared. This information included the pdb code, protein name, X-ray diffraction resolution, an entity of the co-crystalized ligand, organism type, date of depositing the pdb file in the databank, and title of the relevant reported paper. Protein structures from homosapiens sources which had better resolution and were more recently uploaded were preferred. Finally, the 3D structure of the protein set selected from the previous filter was retrieved from the protein data bank in pdb format (16,17).

Protein preparation

The Protein Preparation Wizard (PrepWizard) available in Schrodinger suite 2015 was used to prepare the protein structures in this study (18). This tool allowed us to obtain a reasonable starting structure of the proteins for molecular modeling experiments through a series of computational steps (10). At this stage, the most important modifications applied to the structure of the proteins under study are as follows:

1) Removal of all crystallographic water molecules except for important waters in ligand binding and coordination with metal cofactors.
2) Adding all hydrogen atoms to the structure
3) Creating the lost disulfide bonds in the protein structure. In fact, the goal is to re-establish the missing bonds between the sulfur atoms that are within 3.2 Å of each other.
4) Adjusting the bond orders and formal charges of ligands.
5) Adding and optimizing missing side-chains of residues in the crystal structure of the protein by running Prime side-chain prediction and Prime structure refinement jobs.
6) Filling the missing loops from the SEQRES records in the PDB file of the protein structure and increasing the quality of the resulting loops by running a Prime loop refinement job.

Optimizing the hydrogen-bonding network through predicting protonation states of histidine (His), aspartic acid (Asp), and glutamic acid (Glu), and tautomeric states of histidine; and reorienting hydroxyl and thiol groups, water molecules, amide groups of asparagine (Asn) and glutamine (Gln), and the imidazole ring in histidine (His). These optimizations are necessary because the orientation of hydroxyl/thiol groups, the terminal amide groups in asparagine and glutamine, and the ring of histidine cannot be determined from the X-ray structure.

Performing a restrained energy minimization on the resultant protein structures with an impact refinement module (19), utilizing the OPLS-2005 force field to optimize the geometry and minimize the energy of the protein (20). The minimization was terminated when the energy converged or the root-mean-square deviation (RMSD) reached a maximum cut-off of 0.30 Å.

In fact, after doing the corrections, a notable change in the protein structure occurs, thus, the obtained structure needs to be energetically minimized to take the influence of the structural details of the corrected protein on its total energy into consideration. Finally, the refined protein structures were saved in pdb format.

Model building using the homology modeling method

Prime structure refinement calculations in the previous step are only able to modify the structural defects of amino acids in the side chains. Therefore, to improve the protein structures with missing residues in the main backbone, we used the homology modeling method. In the homology modeling step, we would like to seek an experimentally determined structure possessing a high sequence identity with the target protein. For this purpose, the primary sequence of our target protein was obtained from SWISS-PROT database in FASTA format (Uni-ProtKB /Swiss prot: https://www.uniprot.org/) and the region comprising the missing backbone residues was used for modeling studies. Basic local alignment search tool (BLAST) available at NCBI was employed to find homologous proteins with known structures to be employed as the template in the process of protein homology modeling (21). Accordingly, crystallography structures containing the highest amount of resolved missing residues and with a high sequence identity (> 90%) were selected as rational templates to build problematic regions in the target proteins. MODELLER version 9.25 was used to build homology models (22).

Depending on the situation of missing backbone residues in one or several structural domains, single or multiple sequences alignment was performed between the target and the template proteins, respectively. From the alignments, 3D models containing all non-hydrogen atoms were obtained automatically using the methods implemented in MODELLER (20). From the 1000 model generated with MODELLER for each alignment, the one corresponding to the lowest value of the discrete optimized protein energy (DOPE) was selected as the best model for further evaluation and validation. In this regard, DOPE profiles for individual amino acid residues of models relative to the template were compared using MODELLER. Finally, the stereochemical quality of the final selected models was assessed with respect to the Ramachandran plot generated by PROCHECK web server (23). Moreover, the ProSA webserver was used for further analysis of protein structures and calculation of the Z score parameter of the developed models (24,25).

MD simulation study

The most valid and best models obtained from the protein homology modeling process were further examined for evaluating their thermodynamic behavior and the stability of the protein system using MD simulation studies (26). All MD simulations were carried out by Desmond 4.1 academic version, utilizing Maestro as a graphical interface (27-29). In order to obtain a reasonable protein system as a refined starting point for MD simulation protocol, the candidate models derived from homology modeling studies were initially prepared using the protein preparation wizard workflow as above described (10). In the step of system model building, the prepared complexes were placed into a cubic box filled with water molecules, simulated by the known TIP3P model (30). OPLS_2005 force field was adopted for the MD calculations (20). The system was neutralized using an appropriate number of Na^+/Cl^- counter ions with a fixed salt concentration of 0.15 M which represents the physiological concentration of monovalent ions. Prior to the MD simulation, a series of restrained minimizations and short MD simulations were performed to slowly relax the model system without deviating considerably from the initial protein coordinates. The stages of the pre-relaxation process were: (1) 12-ps simulation in the NVT ensemble (constant number of particles, volume, and temperature: 10 K) restrained with non-hydrogen solute atoms; (2) 12-ps simulation in the NPT ensemble (constant number of particles, pressure, and temperature: 10 K) restrained with non-hydrogen solute atoms; (3) 12-ps simulation in the NPT ensemble (temperature 300 K) restrained with solute non-hydrogen atoms; and (4) 24-ps simulation in the NPT ensemble (temperature 300 K) with no restraints. The temperatures and pressures in the short initial simulations were controlled using Berendsen thermostats and barostats, respectively. Finally, the equilibrated system was simulated for 30 ns at the constant temperature of 300 K and pressure of 1.01325 bar, employing the NPT as ensemble class. RESPA integrator was applied in order to integrate the equations of motion, with an inner time step of 2.0 fs for bonded interactions and non-bonded interactions within the short-range cut-off (31). Nose-Hoover thermostats (32) were used to keep the simulation temperature constant, and the Martyna-Tobias-Klein method (33) was applied to control the pressure. Long-range electrostatic interactions were calculated by the particle-mesh Ewald method (34). The cut-off for van der Waals and short-range electrostatic interactions was set at 9.0 Å. By following this protocol, a single trajectory of 30 ns for each protein system was obtained. The trajectory files were analyzed by simulation event analysis and simulation quality analysis tools provided in the Desmond package. Moreover, the mentioned tools were employed to generate all plots concerning MD simulation analysis included in this study.

RESULTS

Protein dataset and preparation

A dataset of 1001 soluble proteins and their categories and the numbers of the proteins collected in each category are summarized in Table 1. 896 members of the dataset which were corrected and modified are also provided in this table.

Homology modeling

One hundred and five proteins had missing amino acids in their backbone and were subjected to homology modeling. The protein categories and the number of these structures as well as their PDB codes are also provided in Table 2.

Table 1. Protein categories subjected to this study and the number of PDBs in each category.

Protein category	No. of PDBs in the database	No. of PDBs after preparation	Protein category	No. of PDBs in the database	No. of PDBs after preparation
β-Catenin	68	28	P15	183	26
b-Raf Kinase	85	38	P16	35	3
CDK2	127	37	P53	344	167
CDK4	15	4	Telomerase	48	18
Cycline E	71	14	Topoisomerase	271	9
Cyclin D	270	103	TP53	87	36
HDAC	169	25	RTKs	1349	265
HIFs	136	48	WNT	172	75
HSP60	24	0	-	-	-

PDB, Protein data bank; CDK, cyclin-dependent kinase; HDAC, histone deacetylase; HIF, hypoxia-induced factor; HSP60, heat shock protein 60; TP53, tumor protein 53; RTK, receptors tyrosine kinase; WNT, Wingless and Int-1.

Table 2. Protein categories, the number and PDB codes of the protein candidates for homology modeling, and the underlined codes were subjected to this process.

Protein category	No. of proteins in the category	PDB codes
CDK2	4	1jsv, 4ek3, 4fkl, 4nj3
Cyclin D	10	1okv, 1okw, 1ung, 1vym, 2cjm, 2r3i, 2wfy, 4rjf, 6cbi, 6p3w
HDAC	5	1n6j, 1w22, 3ezp, 3mu6, 3rqd
HIFs	2	2ilm, 4b95
P15	1	4b95
P53	33	1gzh, 1h26, 1tsr, 1tup, 2ac0, 2j8z, 2vge, 2vyr, 2wgx, 2wqj, 2wtt, 3d05, 3g03, 3igk, 3kmd, 3lbl, 3sd4, 3ts8, 3v3b, 3vd0, 3zni, 4a63, 4cri, 4hje, 4mzr, 4rg2, 4zfi, 5g4n, 5hob, 6ff9, 6qfm, 6qg8
Telomerase	2	4j19
Topoisomerase	6	1a31, 1a35, 1a36, 1ej9, 1k4t, 4fm9
TP53	13	3igk, 3ts8, 4hje, 4ibu, 4ibv, 4ibw, 4mzr, 4qo1, 5lgy, 5mg7, 6co2, 6ff9, 6i3y
RTKs	26	1agw, 1he7, 1oec, 2itw, 2itz, 2uzx, 2vwz, 2wgj, 2wkm, 2yhv, 2yjr, 2yjs, 3brh, 3f66, 3zbf, 3zfm, 3zxz, 4anl, 4ase, 4f64, 4gt4, 4uwc, 4uxl, 4w4z, 5h3q, 5jr2
WNT	4	1v18, 4a0p, 4uza, 5bpu

PDB, Protein data bank; CDK, cyclin-dependent kinase; HDAC, histone deacetylase; HIF, hypoxia-induced factor; TP53, tumor protein 53; RTK, receptors tyrosine kinase; WNT, Wingless and Int-1.

3D Structure validation and evaluation

Ramachandran plots obtained for each model

The results obtained from the Ramachandran plots of all the modeled proteins are summarized in Table 3. The Ramachandran plot for **1ung** as an example is provided in Fig. 1.

Z-Score plots for the best models obtained

Z-Scores determined by ProSA web server were small negative values located in the dark-blue part of the plot and are provided in Table 3. The Z-score plot of the best homology model of **1ung** is provided in Fig. 2 as an example.

DOPE energy plot for each amino acid in the templates and the obtained model

DOPE energy plots of the template amino acids and the corresponding plot for the best model obtained were similar to the models obtained for all the modeled proteins except **2ilm**. The DOPE energy plot for the best homology model of **1ung** is provided in Fig. 3.

MD simulation results

RMSD plots for the backbone atoms of the modeled proteins

The RMSD plots for the backbone atoms of the modeled proteins are provided in Fig. 4.

RMSF plots for the amino acid residues of the modeled proteins

RMSF plots as an indication of the amino acid residues fluctuations in a protein structure in terms of nanometer in the simulation time are shown in Fig. 5.

Radius of gyration -simulation time plots for the modeled proteins

Time-dependency plots of the radius of gyration for the simulated proteins are provided in Fig. 6.

Table 3. The results were obtained from the Ramachandran plots of all the modeled proteins.

PDB code	Percent of amino acids in the most favored (red) region	Percent of amino acids in the additional allowed (yellow) region	Percent of amino acids in the generously allowed (pale yellow) region	Z-Score
1oec	92.0	7.6	0.4	-7.97
1ung	97.8	2.2	0.0	-6.95
2ilm	76.0	24.0	0.0	-0.31
2itz	93.1	6.2	0.3	-6.58
3rqd	83.5	13.8	1.8	-8.85
3zxz	90.1	8.6	0.7	-8.42
4b95 (Elongin B)	92.0	6.8	1.1	-5.26
4b95 (Elongin C)	94.1	3.5	0.0	-4.33
5g4n	94.2	5.8	0.0	-6.10
5h3q	94.5	4.1	1.5	-6.30
5mg7	85.8	12.4	0.9	-5.99
6cbi	95.8	3.5	0.0	-5.21

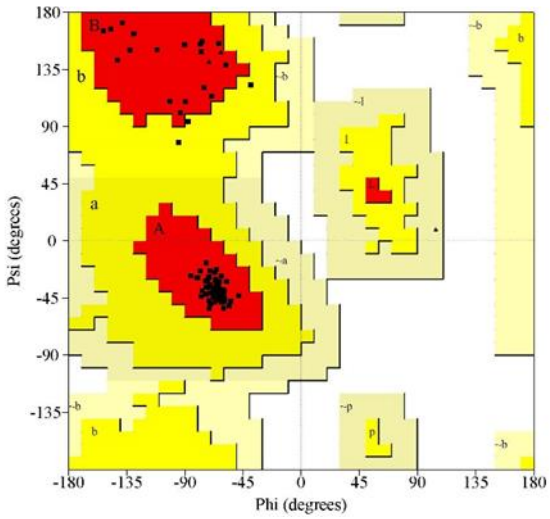

Fig. 1. Ramachandran plot of the best homology model of **1ung**.

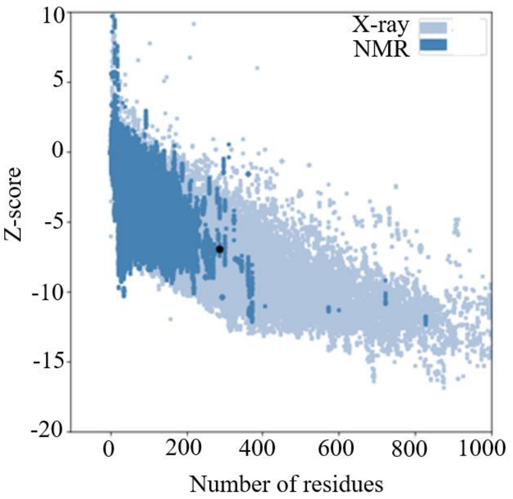

Fig. 2. Z-Score plot of the best homology model of **1ung**.

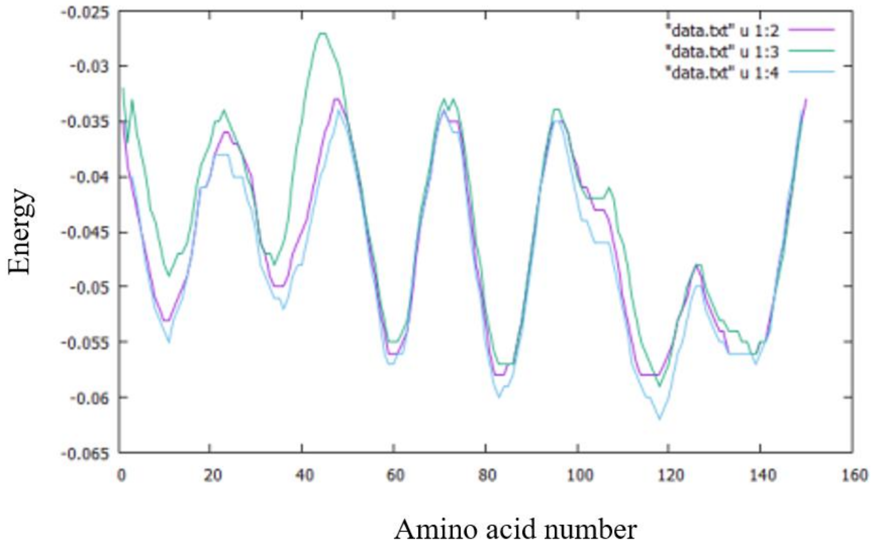

Fig. 3. The discrete optimized protein energy (kcal.mol^{-1}) plot for each amino acid of templates (3: 1h4l) and (4: 3o0g) and the best homology model of **1ung.**

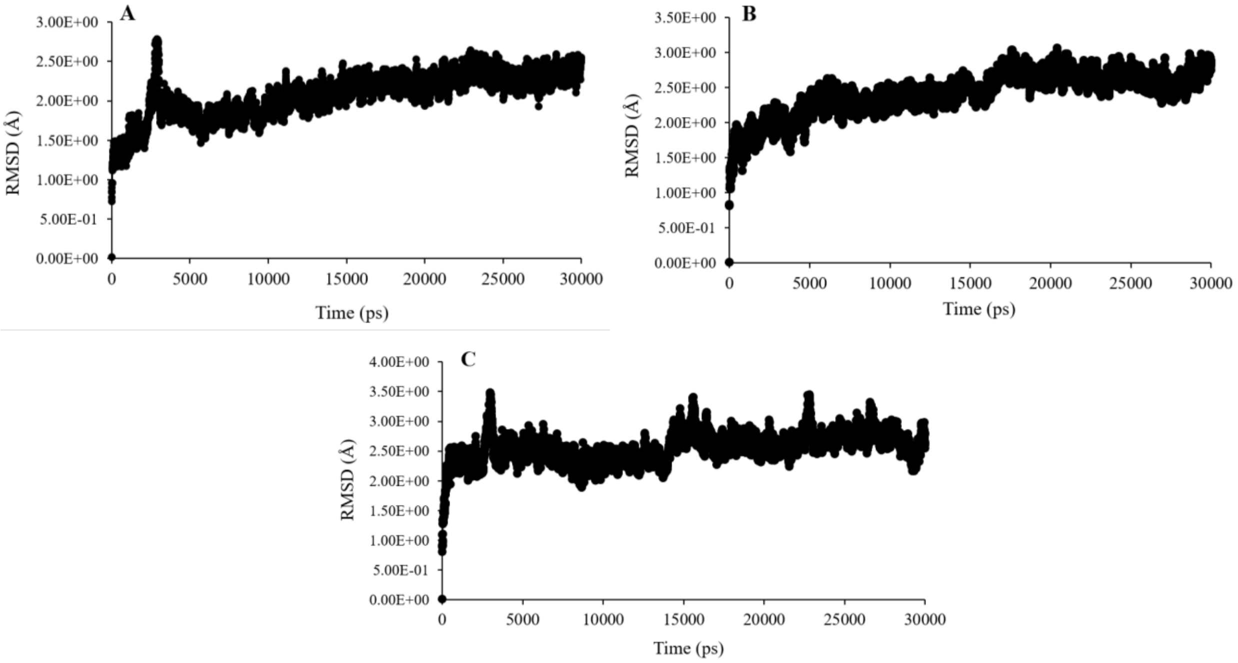

Fig. 4. RMSD of the backbone atoms for three corrected proteins, (A) **1oec**, (B) **5h3q**, and (C) **5g4n**. Changes in RMSD are calculated between the final and initial conformations during 30 ns simulations. RMSD, root-mean-square deviation.

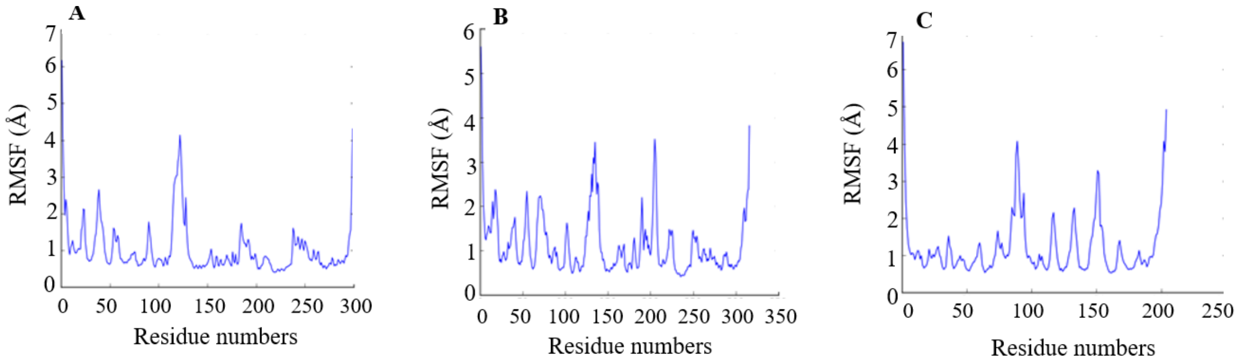

Fig. 5. RMSF of the amino acid residues for three corrected proteins, (A) **1oec**, (B) **5h3q,** and (C) **5g4n**. Changes in RMSF are calculated between the final and initial conformations during 30 ns simulations. RMSF, Root mean square fluctuation.

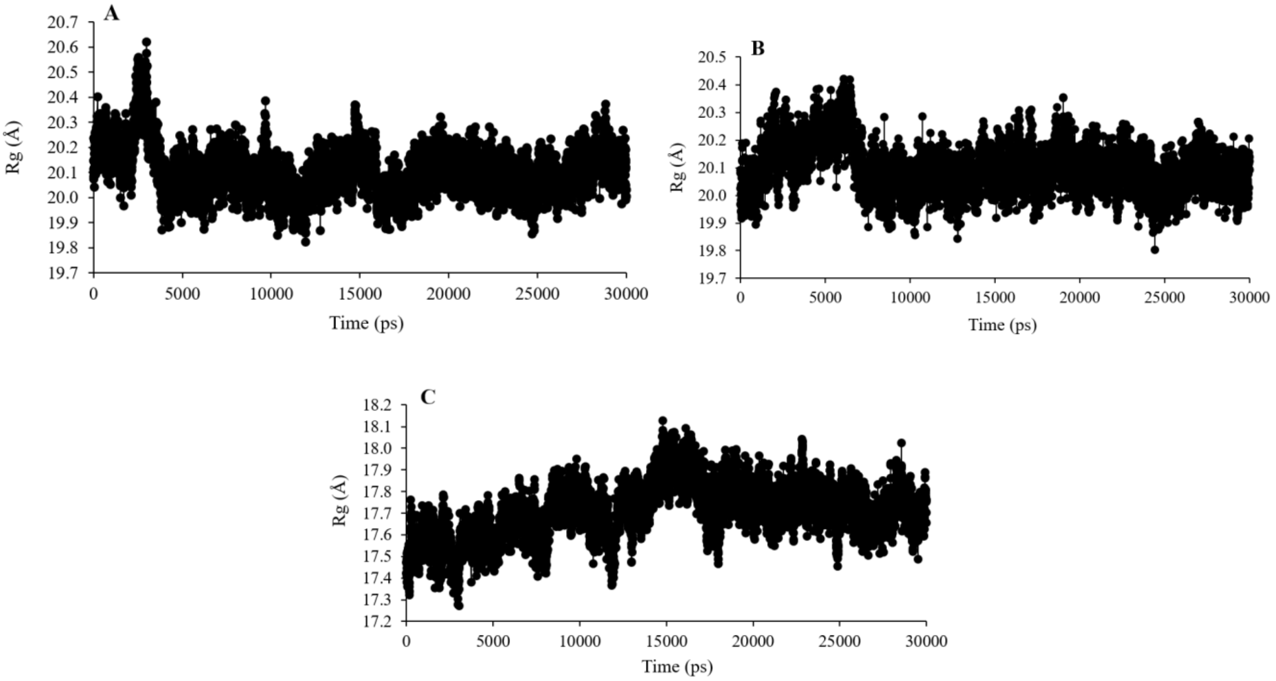

Fig. 6. Rg-simulation time plots for three corrected proteins, (A) **1oec**, (B) **5h3q**, and (C) **5g4n**. Changes in the radius of gyration are calculated between the final and initial conformations during 30 ns simulations. Rg, Radius of gyration.

DISCUSSION

Protein dataset and preparation

In the first step, 3454 soluble proteins of the cell signaling pathways were collected. According to the criteria defined in the methods section, they were filtered to give a dataset of 1001 proteins. All of the members of this database were subjected to corrections in the protein preparation step. In this step, 896 protein structures out of 1001 were modified and corrected all according to the eight well-defined modifications mentioned in the methods section.

One hundred and five protein structures out of 1001 were not perfect after the modifications and corrections in the previous step. In fact, they had something greater than missing in side chains; they bore the missing whole amino acids in some parts of their structure; thus, they were collected for homology model building. Among these candidates, twelve were selected to enter the homology modeling process. The criteria for this selection were the date the PDB code was released, higher resolution, and presence of co-crystallized ligand, and the availability of template/s for homology modeling.

3D Structure validation and evaluation

A thousand models were built for each protein and the best models were chosen based on the DOPE using Ramachandran, Z score, and DOPE plots for every amino acid of the template and the best homology model was obtained.

Evaluation of Ramachandran plots obtained for each model

According to the Ramachandran plots, the most stable modeled protein was the receptor tyrosine kinase protein, 1ung, and the least stable ones were 2ilm 3rqd, and 5mg7. As can be seen in the Ramachandran plot for 1ung, as well as Table 3, 97.8% of the residues are located in the allowed region shown in red color and only 2.2% were in the semi-allowed yellow region.

Evaluation of the Z-Score plots for the best-obtained models

Z-Scores were small negative values located in the dark-blue part of the plot and all modeled proteins except 2ilm were stable according to this parameter.

Evaluation of the DOPE energy plot for each amino acid in the templates and the obtained model

The more the DOPE energy plot of the template amino acids and the corresponding plot for the best model obtained were similar, the more reliable this model was. This was true for the models obtained for all the modeled proteins except 2ilm, a result in accordance with the Ramachandran and Z-score plots.

Analyzing the molecular dynamics (MD) simulation results

Three out of twelve modeled proteins, 1oec, 5h3q, and 5g4n, were subjected to molecular dynamics simulation studies to investigate their structural stability. RMSD, root mean square fluctuation (RMSF), and gyration radius were evaluated to assess the stability and fluctuations of the protein backbone and alpha carbon atoms.

Evaluation of the RMSD plots for the backbone atoms of the modeled proteins

As can be seen in Fig. 4 the average RMSD values for all simulations were zero at the time zero but increased in the simulation period of time until reaching relative stability. The small changes in RMSD values confirmed reaching the structures to thermodynamics equilibrium and the system stability during the 30 ns simulation time.

Evaluation of the RMSF plots for the amino acid residues of the modeled proteins

RMSF is indicative of the amino acid residue fluctuations in a protein structure in terms of a nanometer in the simulation time. The higher RMSF for a protein, in a region not close to the N and C terminals, implies that presumably there is a loop in the protein structure. In other words, the higher value of RMSF shows the higher amount of atomic mobility of the Cα atoms of the protein in the MD simulation, again indicating a loop in the protein structure.

CONCLUSION

Several defects in the x-ray crystal protein structures deposited in protein databases invoke the macromolecule structures sometimes to be unsuitable for structure-based drug design protocols. Mis-differentiation between the NH and O atoms, missing side chains or amino acids, and missing disulfide bonds between cysteine sulfhydryl groups are some of these usual faults. Before starting to work with a crystal structure, researchers have to correct the structural defects to make the structures perfect enough for starting calculations such as molecular modeling, molecular docking, and molecular dynamics simulations. There are just a few databases available on the internet regarding such a subject. In this regard, the database entitled mpstruc "membrane protein of the known 3D structure" (https://blanco.biomol.uci.edu/mpstruc/), developed by the research team of Professor Stephen H. White, at the University of California can be mentioned. In the present research, a small database of the corrected three-dimensional structure files of proteins frequently used in structure-based drug design protocols was introduced. This project covered the modification of some soluble proteins in the cancer signaling pathway. All of the defects mentioned in the introduction were resolved using proper soft-wares and methods. For taking into account the influence of missing amino acids on the protein structure, 12 proteins with missing residues in the backbone were subjected to molecular modeling. Common methods of evaluation confirmed the models according to Ramachandran, z-score, and DOPE energy plots. RMSD, RMSF, and Rg values verified the stability of the models after 30 ns molecular dynamics simulation.

This database will be completed for quite a lot of water-soluble proteins in predicted scheduled steps to be uploaded to the internet.

Acknowledgments

This study was financially supported by the Vice-Chancellery of Research of Isfahan University of Medical Science, Isfahan, Iran through Grant No. 399093.

Conflict of interest statement

The authors declared no conflicts of interest in this study.

Authors' contribution

S. Hatami performed all computational jobs under the supervision of A. Fassihi. H. Sirous Najafabadi, K. Mahnam, and A. Najafipour contributed to the manuscript preparation and revision. The final version of the manuscript was approved by all authors.

REFERENCES

1. Prieto-Martínez FD, López-López E, Juárez-Mercado KE, Medina-Franco JL. Chapter 2-Computational drug design methods-current and future perspectives.in: In silico drug des. Veracruz: 2019; 19-44.
DOI: 10.1016/B978-0-12-816125-8.00002-X.
2. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, *et al*. RCSB protein data bank: enabling biomedical research and drug discovery. Protein Sci. 2020;29(1):52-65.
DOI: 10.1002/pro.3730.
3. Armstrong DR, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, *et al*. PDBe: improved findability of macromolecular structure data in the PDB. Nucleic Acids Res. 2020;48(D1): D335–D343.
DOI: 10.1093/nar/gkz990.
4. Nakamura H. Big data science at AMED-BINDS. Biophys Rev. 2020;12(2):221-224.
DOI: 10.1007/s12551-020-00628-1.
5. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo l, *et al*. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019; 47(D1): D464–D474.
DOI: 10.1093/nar/gky1004.
6. Li Z, Li S, Wei X, Peng X, Zhao Q. Recovering the missing regions in crystal structures from the nuclear magnetic resonance measurement data using matrix completion method. J Comput Biol. 2019;27(5):709-717.
DOI: 10.1089/cmb.2019.0107.
7. Santhosh R, Bankoti N, Padmashri AM, Michael D, Jeyakanthan J, Sekar K. MRPC (missing regions in polypeptide chains): a knowledgebase. J Appl Crystallogr. 2019;52(6):1422-1426.
DOI: 10.1107/s1600576719012330.
8. Ataee MH, Mirhosseini SA, Mirnejad R, Rezaie E, Mahmoodzadeh Hosseini H, Amani J. Design of two immunotoxins based rovalpituzumab antibody against DLL3 receptor; a promising potential opportunity. Res Pharm Sci. 2022;17(4):428-444.
DOI: 10.4103/1735-5362.350243.
9. Razzaghi-Asl N, Mirzayi S, Mahnam K, Adhami V, Sepehri S. *In silico* screening and molecular dynamics simulations toward new human papillomavirus 16 type inhibitors. Res Pharm Sci. 2022;17(2):189-208.
DOI: 10.4103/1735-5362.335177.
10. Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013;27(3):221-234.
DOI: 10.1007/s10822-013-9644-8.
11. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004;3(11):935-949.
DOI: 10.1038/nrd1549.
12. Kellenberger E, Rodrigo J, Muller P, Rognan D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins. 2004;57(2):225-242.
DOI: 10.1002/prot.20149.
13. Feher M, Williams CI. Numerical errors and chaotic behavior in docking simulations. J Chem Inf Model. 2012;52(3):724-738.
DOI: 10.1021/ci200598m.
14. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7-30.
DOI: 10.3322/caac.21590.
15. Farooqi AA, De La Roche M, Djamgoz MB, Siddik ZH. Overview of the oncogenic signaling pathways in colorectal cancer: mechanistic insights. Sem Cancer Biol. 2019;58:65-79.
DOI: 10.1016/j.semcancer.2019.01.001.
16. Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EF, Brice MD, Rodgers JR, *et al*. The Protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977;112(3):535-542.
DOI: 10.1016/s0022-2836(77)80200-3.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al*. The protein data bank. Nucleic acids Res. 2000;28(1):235-242.
DOI: 10.1093/nar/28.1.235.
18. Protein Preparation Wizard 2015, -1; Epik version 2.4, Schrödinger, LLC, New York, NY, 2015; Impact version 5.9, Schrödinger, LLC, New York, NY, 2015; Prime version 3.2, Schrödinger, LLC, New York, NY, 2015.
19. Impact, Impact Version 5.9. Schrödinger, LLC, New York, NY, 2015.
20. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc. 1996;118(45):11225-11236.
DOI: 10.1021/ja9621760.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped blast and PSI-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-3402.
DOI: 10.1093/nar/25.17.3389.

22. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol. 1993:234(3);779-815.
DOI: 10.1006/jmbi.1993.1626.
23. Laskowski RA, Macarthur MW, Moss DS, Thornton J. Procheck: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993;26(2):283-291.
DOI: 10.1107/S0021889892009944.
24. 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):W407-W410.
DOI: 10.1093/nar/gkm290.
25. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993;17(4):355-362.
DOI: 10.1002/prot.340170404.
26. Adcock SA, McCammon JA. Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev. 2006;106(5): 1589-1615.
DOI: 10.1021/cr040426m.
27. Maestro, In Maestro Version 10.1; Schrödinger, LLC: New York, NY, USA, 2015.
28. Desmond Molecular Dynamics System, version 4.1,D. E. Shaw Research, 2015. Maestro-Desmond Interoperability Tools, Version 4.1. Schrödinger, New York, NY, 2015.
29. Bowers KJ, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, *et al*. Scalable algorithms for molecular dynamics simulations on commodity clusters. Proceedings of the 2006 ACM/IEEE conference on Supercomputing, Florida: IEEE; 2006. pp. 84-96.
DOI: 10.1109/SC.2006.54.
30. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J chem phys. 1983;79(2):926-935.
DOI: 10.1063/1.445869.
31. Humphreys DD, Friesner RA, Berne BJ. A multiple-time-step molecular dynamics algorithm for macromolecules. J Phys Chem. 1994;98(27):6885-6892.
DOI: 10.1021/j100078a035.
32. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985;31(3):1695-1697.
DOI: 10.1103/PhysRevA.31.1695.
33. Martyna GJ, Tobias DJ, Klein ML. Constant pressure molecular dynamics algorithms. J Chem Phys. 1994;101(5):4177-4189.
DOI: 10.1063/1.467468.
34. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995;103(19):8577-8593.
DOI: 10.1063/1.470117.