Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): ------- (2025)
Copyright © Pakistan Academy of Sciences
ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)693

Pakistan Academy of Sciences

Research Article

Structure Prediction of the *Bombyx mori* Sericin 4 Protein

Khushnudbek Eshchanov[*], **Dono Babadjanova, and Mukhabbat Baltaeva**

Department of Chemistry, Urgench State University, Urgench, Uzbekistan

Abstract: Natural silk (*Bombyx mori*) has been found to contain sericin 1, sericin 2, sericin 3, and sericin 4 proteins. The sequence of amino acid residues in them has also been well studied. However, there is little information on the molecular structure of sericin 4. We conducted studies on the prediction of the sericin 4 molecule's structure using the AlphaFold3 and YASARA computational servers. Molecular dynamics simulations were performed in aqueous solution to evaluate the stability and determine the most favourable conformation of the predicted sericin 4 structure. We mainly used the ProSA-web, Ramachandran Z and Molprobity score to evaluate the predicted structure of sericin 4, and the reliability of the predicted model was determined. The predicted molecular structure serves as a preliminary, yet robust, model of sericin 4.

Keywords: Sericin 4, Silk, Ramachandran Z-Score, Minimum Energy, Solubility, Structure.

1. INTRODUCTION

Proteins extracted from natural silk raw materials are considered as important biomaterials that are the focus of current research. Silk sericin protein is important due to its water solubility, antioxidant properties, biodegradability, and suitability for the preparation of biomaterials for medicine [1-3]. Sericin is often recognised as an "adhesive" protein, enveloping the silk fibroin of *Bombyx mori* and constituting 20–30% of its total mass [4]. In recent years, sericin has been widely employed in nanocomposites, hydrogels, and tissue engineering (for instance, in skin regeneration and wound healing), yielding positive outcomes in its clinical trials [5, 6]. To evaluate and consider the potential uses of sericin, knowledge of its properties, structure, and composition is required [7, 8].

Sericin is a globular protein characterised by the presence of random coils and β-sheet structures. Several external factors, including temperature, humidity, and mechanical stress, can influence the transition of sericin from a random-coil conformation to a β-sheet arrangement. Sericin is highly soluble in water at temperatures of 50 °C and above [9]. This structural transition is thermodynamically linked to a reduction in entropy, and parameters such as pH and ionic strength further affect the kinetics of gel formation [10]. For example, at physiological pH (pH 7), the gelation process can proceed two to three times faster. In contrast, at lower temperatures, the solubility of sericin diminishes, promoting the conversion of random coils into β-sheets and consequently leading to gel formation [11]. Moreover, it has been demonstrated that higher sericin concentrations accelerate the gelation process [12]. Sericin is a hydrophilic protein, distinguished by a high proportion of free hydroxyl (-OH), carboxyl (C=O), and other polar functional groups within its amino-acid residues [13]. Its amino acid composition is dominated by serine (Ser, 37%), glycine (Gly, 16%), and aspartic acid (Asp, 15%), which ensures its high hydrophilicity [14].

It has been found that there are 4 different types of sericin 1, sericin 2, sericin 3, and sericin 4 proteins in *Bombyx mori* silk fiber [4]. These sericin proteins in silk fiber glue together two fibroin fibers. The structure and composition (amino acid sequence) of sericin 1, sericin 2, as well as sericin 3 proteins have been well studied by previous researchers [15, 16]. Komatsu [17] determined the amounts of sericin 1, sericin 2, sericin 3, and sericin 4 proteins in an aqueous solution of sericin

Received: July 2025; Revised: November 2025; Accepted: December 2025

[*] Corresponding Author: Khushnudbek Eshchanov <xeshchanov77@gmail.com>

extracted from *Bombyx mori* cocoons, and showed that the amount of sericin 4 was 3.1%. The low content of sericin 4 indicates its specific role in interaction with fibroin, it is primarily located in the inner layers and contributes to mechanical strength. This protein serves as a protective and binding component that surrounds the fibroin filaments. Therefore, determining the molecular structure of sericin 4 provides not only insight into its unique physicochemical properties but also a deeper understanding of the surface behaviour of silk-based biomaterials.

The structural uniqueness of sericin 4 is reflected in its amino acid composition and polypeptide chain arrangement. It is rich in polar amino acids such as serine, asparagine, and threonine, which impart a highly hydrophilic character to the protein. As a result, sericin 4 readily interacts with water molecules, thereby contributing to the surface moisture of silk. This property enhances the biocompatibility of silk materials and is particularly important for their biomedical applications, such as in wound dressings, drug delivery systems, and biopolymer films [18].

Information about sericin proteins is also included in the Uniprot and Swiss databases. The Uniprot database accurately describes the 3D molecular structures of sericin proteins and their amino acid sequences [19, 20]. Many scientific publications have been published that fully confirm this information. However, the 3D molecular structure of the sericin 4 protein is poorly understood. It should also be noted that successful work has been carried out to determine the amino acid sequence of sericin 4 [21]. However, the molecular structure of the sericin 4 molecule remains elusive. To some extent, it is possible to predict the formation of the sericin 4 protein to solve this problem. Using the latest AlphaFold3 and RoseTTAFold models, it is possible to predict the approximate 3D structure of sericin 4, which may reveal its β-sheet richness (45%) and potential disulphide bridges [22].

Protein structure prediction relies on the amino acid sequence. The secondary and tertiary structures are inferred from the primary structure. It should be noted, however, that the predicted structure may differ slightly from the protein's actual conformation [23]. The protein chain can adopt numerous conformations due to rotation around the φ and ψ torsion angles at the Cα atom. This conformational freedom contributes to variations in the three-dimensional architecture of proteins. Peptide bonds within the chain are polar, containing carbonyl and -NH- groups that are capable of forming hydrogen bonds. As a result, these groups interact within the protein and play a crucial role in stabilising its structure. Glycine holds a distinctive position in protein architecture, as its minimal side chain grants it increased local flexibility. In contrast, cysteine residues may react with one another to form disulfide bonds, creating cross-links that reinforce the overall stability of the protein. Protein structure is commonly described in terms of secondary structural elements, such as α-helices and β-sheets. Within these motifs, regular hydrogen-bonding patterns arise between the -NH- and C=O groups of neighbouring amino acids, and the residues typically possess similar φ and ψ torsion angles [24].

The development of secondary structural elements enables the hydrogen-bonding potential of peptide bonds to be effectively fulfilled. These secondary structures may be densely packed within the hydrophobic core of a protein, although they may also be found on the surface where the environment is polar. Each amino-acid side chain occupies a finite volume and can engage in only a limited range of interactions with neighbouring residues; such steric and interaction constraints must be carefully considered in molecular modelling and sequence alignment studies [25]. The Ramachandran plot is employed to identify the energetically allowed regions for φ and ψ torsion angles, thereby demonstrating the thermodynamic favourability of β-sheet formation in Sericin 4.

Protein structures can be experimentally identified using methods such as X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance (NMR) spectroscopy. However, these approaches are both costly and time-consuming. Over the past six decades, experimental efforts have resolved the structures of approximately 170000 proteins, despite the fact that more than 200 million proteins are known across all forms of life. By 2025, the AlphaFold database had predicted structures for over 214 million proteins, yet certain rare proteins, including sericin 4, have not been fully verified experimentally. Throughout recent decades, numerous computational strategies have been developed to infer three-dimensional protein structures directly from amino-acid sequences. In the most successful cases, homology-based modelling grounded in molecular evolution has achieved accuracy approaching that of experimental methods, such as NMR spectroscopy [26]. Precise protein-structure prediction holds major importance in fields such as drug discovery and biotechnology [27-29].

Protein structure prediction represents one of the central objectives of computational biology and is closely related to the resolution of the Levinthal paradox. Levinthal's paradox is a conceptual experiment in the context of protein-folding studies, highlighting that protein folding involves identifying the most energetically stable conformation. Exhaustively searching all possible structural conformations to locate the lowest-energy state would be computationally impractical. Yet, in nature, proteins fold extremely rapidly - even when adopting highly complex topologies - indicating that folding proceeds through a rugged energy landscape that guides the molecule efficiently towards a stable configuration [30]. Levinthal also demonstrated that, in cases where the global minimum energy state is not kinetically accessible, proteins may adopt a metastable conformation with slightly higher energy [31]. The most effective approaches in structural bioinformatics tend to be those that build upon existing biological and structural knowledge, rather than attempting to model protein folding entirely from first principles.

When predicting a protein structure or evaluating the quality of a homology model, it is highly beneficial to first examine a large number of experimentally determined structures to gain an understanding of what the actual protein may look like. This comparative insight facilitates a more accurate assessment of the model's reliability and structural validity. Many servers have been created for protein structure prediction. The AlphaFold3 server occupies a special place in protein structure prediction and is the leading server. AlphaFold3 is not limited to single-chain proteins, as it can also predict the structures of RNA, DNK, post-translational modifications, and protein complexes with selected ligands and ions. The AlphaFold3 server allows for structure prediction of proteins consisting of sequences of up to 5000 amino acid residues [32-34].

The Ramachandran Z-score is also regarded as a reliable indicator for the overall assessment of protein structures. Hooft *et al*. introduced this numerical measure, known as the Ramachandran Z-score (Rama-Z), to characterise the distribution of φ and ψ torsion angles in the Ramachandran plot. Its primary significance lies in its ability to indicate the structural credibility of newly determined protein models. The Rama-Z score functions as a global metric, offering an overall evaluation of model quality, although it does not identify local deviations in main-chain geometry. In addition to the single global score, separate Rama-Z values are also computed for β-strands, α-helices, and loop regions. Nevertheless, the global Rama-Z score remains the most informative measure for general structural validation. The value of the Rama-Z score correlates with the proportion of residues that fall within the favourable regions of the Ramachandran plot. Analyses of models resolved at 1.2–5 Å resolution demonstrated that 28% exhibited Rama-Z < -2, 14% had Rama-Z < -3, 0.19% displayed Rama-Z > 2, and only 0.01% had Rama-Z > 3. Based on these observations, a protein structure is considered acceptable when its Rama-Z score lies within the range -3 to 3 [34].

We attempted to demonstrate the 3D molecular structure of sericin 4 based on the latest information on its amino acid sequence, and studies have been conducted. In this work, the potential conformations of sericin 4 are analysed using AlphaFold3 and molecular dynamics (MD) simulations, which may reveal its novel applications as a biomaterial.

2. MATERIALS AND METHODS

Using the AlphaFold3 server, CIF and JSON files were generated (by entering the amino acid residue sequences of sericin 4) for five distinct models of the predicted protein structure. However, the generated models contain structural errors. The model with the fewest errors was identified using dedicated evaluation servers. ProSA-web and Ramachandran Z-scores were employed to provide an overall assessment of the protein structures. The ProSA-web server determines the similarity of protein structures to those characterised by X-ray and NMR analyses; low similarity may indicate the presence of structural errors [25, 26]. The sericin 4 structure was evaluated using MolProbity, one of the most reliable validation tools available. To achieve favourable validation metrics, defects in the protein structure were minimised using the YASARA minimization server [35]. This server performs an energy minimisation using the YASARA force field. Iterative refinement of the sericin 4 molecular model was performed via this server to optimise the structure. Subsequently, the stability of the sericin 4 model in aqueous solution was investigated through molecular dynamics (MD) simulations. Computations were conducted using the OPLS-AA/L force field and the SPCE water model within the GROMACS MD package, as implemented in the BioExcel Building Blocks Workflows platform. The reliability of the optimised model was reassessed using MolProbity.

3. RESULTS AND DISCUSSION

The presence of four sericin proteins in *Bombyx mori* silk has been reported in the literature [4, 17]. UniProt, Swiss-Prot, and other protein databases contain extensive information on the composition, structure, and other properties of sericin 1, sericin 2, and sericin 3. These databases do not contain information about sericin 4. However, studies have been conducted to determine the structure of sericin 4, and positive results have been reported. Ping Zhao et al. have published research on the sequence of amino acid residues in the sericin 4 molecule. They analysed sericin 4 in terms of its chain segments based on the amino acid residue sequence [20]. This study did not, however, provide information on the complete structure of sericin 4.

The three-dimensional structure of Sericin 4 was predicted using the AlphaFold server based on its amino acid sequence, and comparative analyses were performed to select the most reliable structural model. The sericin 4 protein consists of 2296 amino acid residues, with the largest proportions being Lys (9.7%), Thr (9.4%), Ser (9.4%), Glu (8.9%), and Gly (7.4%). The theoretically calculated isoelectric point (pI) is 6.25. As shown in Figure 1, the following structural models were predicted by the AlphaFold server based on the amino acid residue sequence of sericin 4.

Calculations were carried out using the ProSA-web server to evaluate which of the derived sericin 4 molecular models was the most reliable. ProSA-web determines an overall quality score for the submitted structure. If this score falls outside the range typical of native proteins, the structure may contain errors. The local quality score diagram highlights problematic regions within the model. A three-dimensional molecular representation

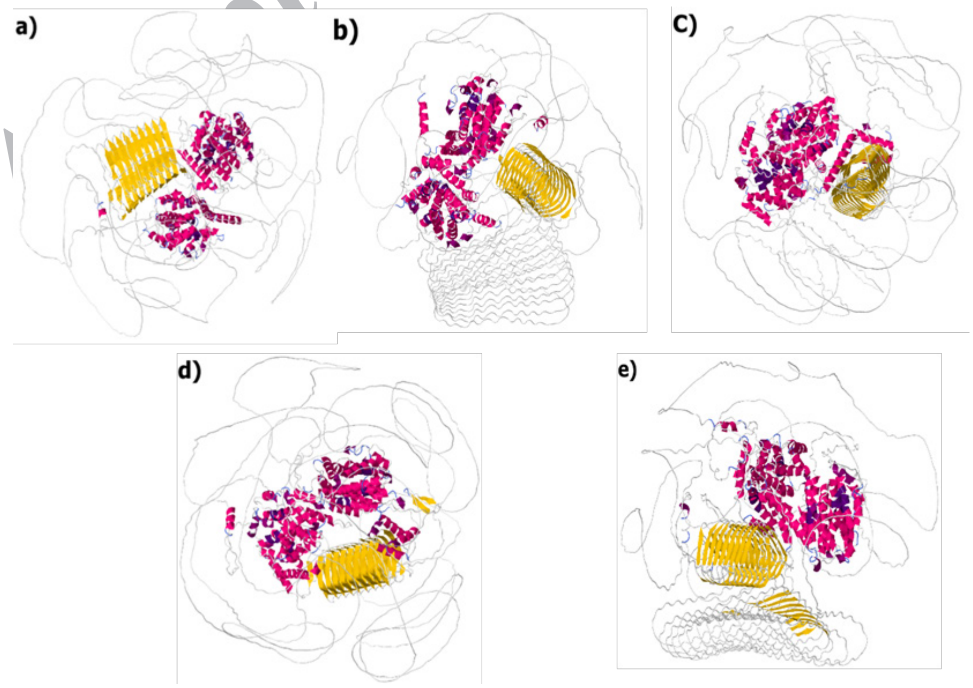

Fig. 1. Models of the sericin 4 molecule created using the AlphaFold3 computational server (Five different molecular models: (a) Compact β-barrel-rich globular model, (b) Extended loop-dominant unfolded-like model, (c) Intermediate partially folded β-sheet model, (d) Globular model with central β-barrel core', and (e) Elongated multi-domain flexible model).

can also be generated to aid in the identification of such areas. ProSA-web is applicable to both low-resolution structures and approximate models obtained during the early stages of structural determination.

The Z-score reflects the overall quality of the model. Its value is displayed on a graph containing the Z-scores of all experimentally determined protein chains, with those derived from different experimental techniques (X-ray and NMR) indicated in distinct colours [25, 26]. The Z-score of a protein is defined as the energy separation between the local fold and the mean value of an ensemble of misfolded folds, expressed in units of the ensemble's standard deviation. It has been reported that calculated Z-scores are generally smaller than experimental values [32, 33].

The results showing the Z-scores for the sericin 4 models generated by the AlphaFold server, and indicating chain segments with relatively higher energy, are presented in Figure 2. The Z-scores for models "a", "b", "c", "d", and "e" of sericin 4 were 0.53, -7.55, -1.52, -6.3, and -8.51, respectively. Examination of these values reveals that the lowest score (-8.51) corresponds to the "e" model structure.

In Figure 2(I-V), illustrating problematic or erroneous regions of the structures, positive values indicate faulty areas. The single-residue energy diagram typically exhibits large fluctuations and is therefore of limited use in model assessment. The greater the number of lines representing negative energy regions, the fewer the structural defects, and thus the more reliable the model. Based on these results, the "e" model of sericin 4 (Z-score -8.51) can be regarded as the most reliable structure.

The sericin 4 models were also evaluated using the global Ramachandran Z score (Rama-Z). The results obtained are presented in Table 1.

The Rama-Z score serves as a global indicator for assessing the overall quality of a protein model and does not provide information on local backbone alignment issues. It is important to highlight that, in addition to the single global Rama-Z value, individual Rama-Z scores are also determined for coils, helices, and β-sheets. A model is generally considered accurate and reliable when its Rama-Z score falls within the range of -3 to 3 [34]. Based on the structural evaluation of sericin 4, it can be observed that the Rama-Z score for the "e" model lies relatively close to -3.

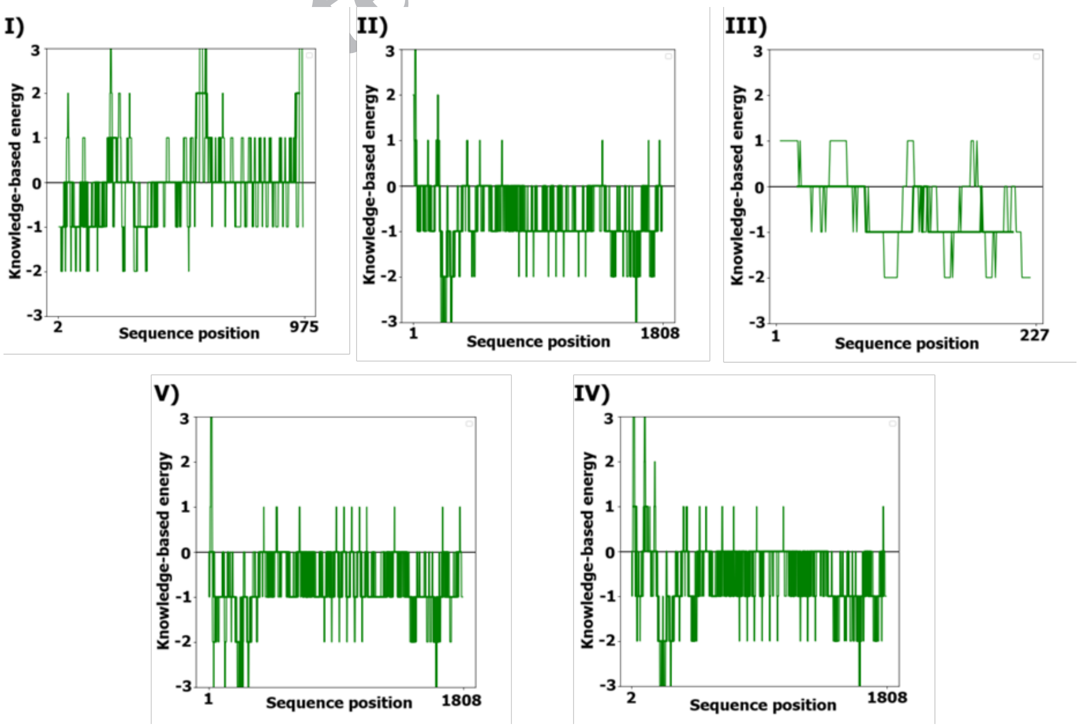

Fig. 2. Diagrams showing high-energy chain segments in models of the sericin 4 molecule: (I) a-Compact β-barrel-rich globular model, (II) b-Extended loop-dominant unfolded-like model, (III) c-Intermediate partially folded β-sheet model, (IV) d-Globular model with central β-barrel core', and (V) e-Elongated multi-domain flexible model.

Table 1. Ramachandran Z score values of sericin 4 molecular models.

Molecular model	Ramachandran Z-score	Side-chain Z-score
a) Compact β-barrel-rich globular model	-6.08	-2.27 ± 0.22
b) Extended loop-dominant unfolded-like model	-4.52	-1.14 ± 0.22
c) Intermediate partially folded β-sheet model	-5.31	-1.80 ± 0.22
d) Globular model with central β-barrel core'	-5.30	-1.91 ± 0.21
e) Elongated multi-domain flexible model	-4.51	-0.89 ± 0.22

The YASARA minimisation server was used to correct energetically unfavourable regions in the "e" model chain of the sericin 4 molecule and to improve its geometry. The YASARA minimisation server is invaluable in protein structure determination, as it provides a realistic impression of the protein's native conformation and demonstrates how to assess the accuracy of the refined model [35]. Using the YASARA minimisation server, the energy of the "e" model of sericin 4 was reduced to its minimum state (Figure 3).

The model was energy-minimised using the YASARA minimisation server for 57 cycles. The Rama-Z score was again used to evaluate the overall structure of the energy-minimised model. The model exhibiting the best Rama-Z score of -2.72 and a minimum energy value of -1069996.7 kJ/mol is presented in Figures 3 and 4. However, according to the MolProbity analysis, among all energy-minimised structures, the model obtained after 51 optimisation cycles in the YASARA program demonstrated the highest quality score, indicating the lowest level of structural errors (Figure 5).

MolProbity is a widely recognised platform for evaluating the geometrical and all-atom quality of three-dimensional macromolecular models, including proteins, nucleic acids, and ligands. It provides detailed validation metrics such as clash scores, Ramachandran plot and rotamer outliers, Cβ deviations, and the overall MolProbity score [36]. The model optimised 51 times achieved a MolProbity score of 1.25, suggesting a high-quality and well-refined structure. The summarised validation results are presented in Table 2.

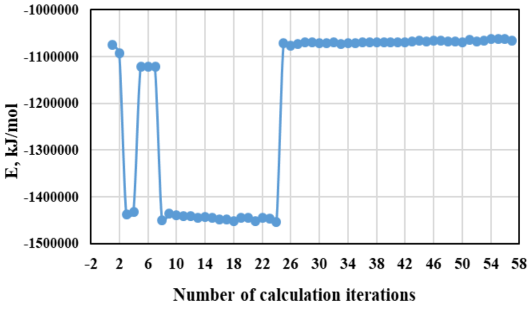

Fig. 3. Minimum energy results of the "e" model of sericin 4 in iterative calculations using the YASARA minimisation server.

MolProbity analysis reveals that the protein structure is of high quality: Clashscore 0.45 (99th percentile) and MolProbity score 1.25 (99th percentile) - placing it within the top 1% of PDB entries. Steric clashes and overall geometry are excellent. Ramachandran favoured 88.49% (<98%) - slightly low, but outliers (0.96%) remain within acceptable limits. CaBLAM (6.1%) and CA outliers (3.14%) are acceptable for lower-resolution structures.

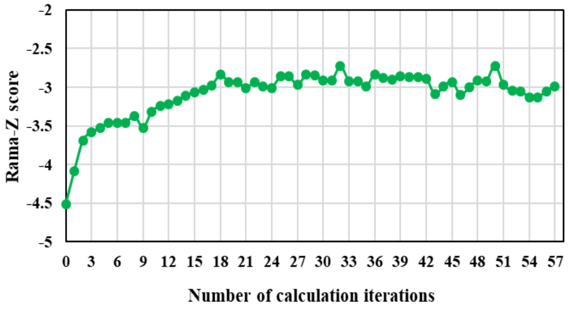

Fig. 4. Rama-Z scores of "e" model sericin 4 that were re-minimised 57 times in the YASARA minimisation server.

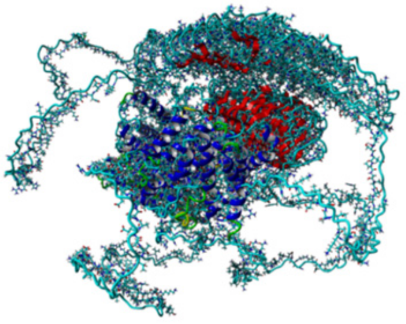

Fig. 5. Energy minimised model of sericin 4 by the YASARA minimisation server.

To mitigate structural inconsistencies observed in the sericin 4 model, the Rosetta Relax refinement was applied [37]. This approach resulted in a notable improvement in the overall structural quality, as evidenced by the evaluation metrics presented in Table 3.

Molecular dynamics (MD) simulation is one of the most powerful computational techniques for investigating the structural and functional properties of proteins at the atomic level. Unlike static crystallographic structures, MD provides a realistic description of the time-dependent dynamic

Table 2. Molprobity analysis of Sericin 4 molecular structures optimised 51 times using YASARA minimisation server.

Clashscore, all atoms	**0.45**	**99th percentile*(N=1784, all resolutions)**
Poor rotamers	0.90%	Goal: <0.3%
Favored rotamers	96.65%	Goal: >98%
Ramachandran outliers	0.96%	Goal: <0.05%
Ramachandran favored	88.49%	Goal: >98%
Rama distribution Z-score	-2.24 ± 0.15	Goal: abs(Z score) < 2
MolProbity score^	**1.25**	**99th percentile* (N=27675, 0Å - 99Å)**
Cβ deviations >0.25Å	0.19%	Goal: 0
Bad bonds:	0.25%	Goal: 0%
Bad angles:	0.39%	Goal: <0.1%
Cis Prolines:	8.70%	Expected: ≤1 per chain, or ≤5%
Twisted Peptides:	0.04%	Goal: 0
CaBLAM outliers	6.1%	Goal: <1.0%
CA Geometry outliers	3.14%	Goal: <0.5%
Chiral volume outliers	0/2720	
Waters with clashes	0.00%	See UnDowser table for details

Table 3. MolProbity analysis of sericin 4 structures refined with Rosetta Relax.

Clashscore, all atoms:	**1.96**	**99th percentile*(N=1784, all resolutions)**
Poor rotamers	0.00%	Goal: <0.3%
Favored rotamers	99.95%	Goal: >98%
Ramachandran outliers	1.05%	Goal: <0.05%
Ramachandran favored	94.07%	Goal: >98%
Rama distribution Z-score	-0.78 ± 0.16	Goal: abs(Z score) < 2
MolProbity score^	**1.36**	**99th percentile* (N=27675, 0Å - 99Å)**
Cβ deviations >0.25Å	0.00%	Goal: 0
Bad bonds:	0.07%	Goal: 0%
Bad angles:	0.13%	Goal: <0.1%
Cis Prolines:	8.70%	Expected: ≤1 per chain, or ≤5%
Twisted Peptides:	0.00%	Goal: 0
CaBLAM outliers	5.4%	Goal: <1.0%
CA Geometry outliers	2.49%	Goal: <0.5%
Chiral volume outliers	0/2720	
Waters with clashes	0.00%	See UnDowser table for details

behaviour of biomolecules. Through MD, the motion of each atom within the protein is computed based on Newtonian mechanics, allowing the exploration of energetically favourable conformations, internal flexibility, and vibrational motions within the system. By evaluating the stability of a protein structure, MD simulation helps to identify the lowest potential energy conformation, which often corresponds to its biologically active form. Therefore, it significantly contributes to energy minimisation and a more accurate representation of the native structural state. Moreover, the simulation enables the analysis of a protein's flexibility, its response to environmental conditions such as temperature and pH, and its interaction mechanisms with ligands or substrates.

Additionally, molecular dynamics complements experimental methods such as X-ray crystallography and NMR spectroscopy by providing time-resolved atomic-level information. The combination of MD data with experimental results allows researchers to construct a more complete and realistic molecular model that explains the functional mechanism, stability, and conformational transitions of the protein. Based on this data, calculations were performed using the MD method for the sericin 4 molecule.

Molecular dynamics (MD) simulations were performed on the BioExcel Building Blocks Workflows platform using the GROMACS MD package with the OPLS-AA/L force field and the SPCE water model [38]. In the simulation setup, a single protein molecule was solvated with 10000 water molecules, 956 Na^+ ions, and 910 Cl^- ions. The net charge of the protein was -46. The simulation lasted for 100 nanoseconds (ns), and the molecular structure was optimised.

The RMSD (Root mean square deviation) graph shows how the shape of the molecule changes over time (Figure 6). In the graph, the RMSD increases from 0 ps to 500 ps and stabilises around 0.4 nm. This indicates that the molecule initially underwent a rapid conformational adjustment (adaptation phase) and subsequently reached a stable state. The RMSD value suggests that the molecule has deviated to some extent from its initial conformation; however, this does not imply instability. Rather, it is associated with the molecule's transition to a new, energetically favourable conformation. Structural stability was achieved after approximately 200-300 ps, and the system remained stable overall.

The radius of gyration (Rg) was also analysed, and the corresponding results are shown in the graph. Rg reflects the compactness or degree of expansion of the molecule. The overall Rg value remained nearly constant at around 4.8 nm. The RgX, RgY, and RgZ values along the three axes also showed very little fluctuation. This indicates that the molecule maintained its general shape, meaning that it neither compressed nor expanded noticeably. Therefore, compactness and structural stability were preserved throughout the entire simulation. Conformational changes were minimal, and the molecule remained in a stable configuration (Figure 7).

The energetic states of sericin 4 were assessed based on the "GROMACS Energies" plot, which shows the potential and total energy (Figure 8). Both energy values remained nearly constant over 500 ps, with only minor fluctuations. The potential energy stabilised around $-16 \cdot 10^6$ kJ/mol, and the

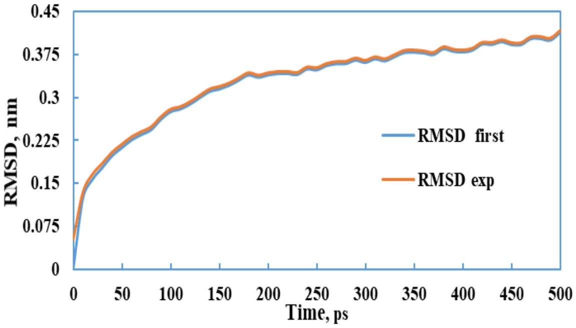

Fig. 6. Root mean square deviation plot of sericin 4 molecule.

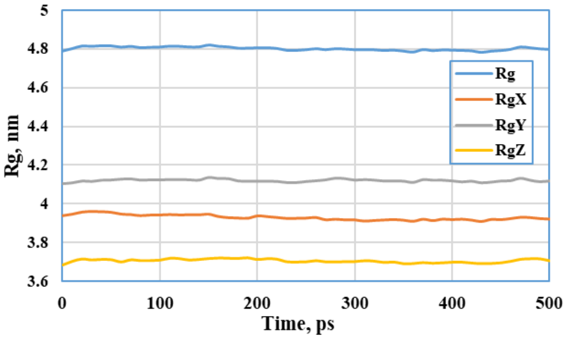

Fig. 7. Stability analysis of sericin 4 based on radius of gyration (Rg).

total energy around -13.5·10^6 kJ/mol. The very small fluctuations indicate that the system reached thermal equilibrium. No significant variations or signs of instability were observed in the results (Figure 9).

The molecular weight, isoelectric point, and other parameters of sericin 4 were determined using the ExPASy (ProtParam) server. The results are presented in Table 4. This server can help to accurately calculate many protein parameters [39-41].

The CamSolpH computational server was used to theoretically study the dependence of the solubility of the improved model of sericin 4 on the pH value of the medium in the YASARA minimization server. CamSolpH provides a solubility profile, where regions with a score greater than 1 indicate highly soluble regions and regions with a score less than -1 indicate poorly soluble regions. The entire sequence is given an overall solubility score. This score can be used to rank different protein variants with high accuracy according to their solubility [42].

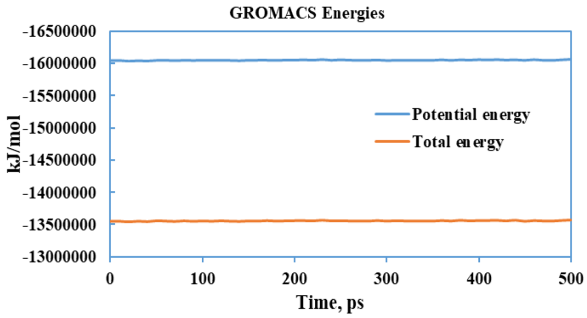

Fig. 8. Potential and total energy stability of the sericin 4 protein during MD simulation.

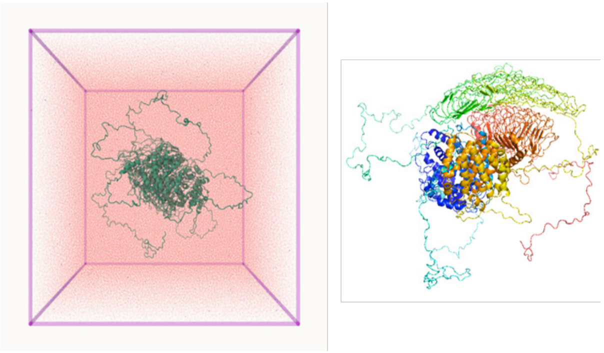

Fig. 9. Conformational state of the Sericin 4 molecule resulting from molecular dynamics simulation.

If we look at Figure 10, the CamSolpH score is greater than 1 in the range of pH values in the solvent (water) medium from 1 to 14. This value theoretically confirms that sericin 4 has good solubility. When comparing the relative solubility at different pH values, it can be seen that the solubility is lowest at pH = 10. It can be assumed that the solubility of sericin 4 is highest in solvents with a pH value of up to 4. However, an increase in solubility can be observed in solvents with a pH value higher than 10.

Table 4. Some calculated parameters of sericin 4.

Molecular model	Parameters
Amino acid number	2296
Molecular weight	254369.63 Da
Isoelectric point	6.25
Extinction coefficients (in water, 280 nm)	175395 $M^{-1}·cm^{-1}$
The instability index	43.88

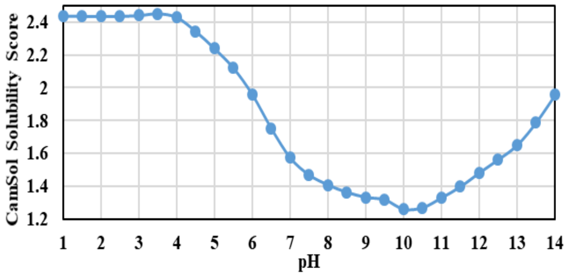

Fig. 10. Solubility index of sericin 4 in solvents (water) with different pH values.

4. CONCLUSIONS

In this study, a comprehensive computational investigation was carried out to predict and analyse the structural and dynamic properties of the sericin 4 protein from *Bombyx mori*. Since no experimental data are available in protein databases, structural prediction was initially performed using the AlphaFold server, yielding five possible molecular conformations. Comparative evaluation through ProSA-web analysis identified the "e" model (elongated multi-domain flexible model) as the most reliable structure, with the lowest Z-score (-8.51). Further refinement using the YASARA minimisation server reduced the overall potential energy of the structure to its minimum state and improved its geometry. Furthermore, refinement with the Rosetta Relax resulted in an additional improvement of the sericin 4 structure. MolProbity validation confirmed the high quality of the optimised model (MolProbity score 1.36, Clashscore 1.96, 99th percentile, Rama distribution Z-score -0.78 ± 0.16, favored rotamers 99.95%), suggesting that the refined model accurately represents the likely native conformation of sericin 4. Molecular dynamics (MD) simulations performed with GROMACS (OPLS-AA/L force field and the SPCE water model) demonstrated the structural stability of the sericin 4 molecule over a 100 ns trajectory. The RMSD and radius of gyration (Rg) analyses indicated that the protein achieved a stable conformational equilibrium after approximately 200-300 ps, maintaining compactness and structural integrity throughout the simulation. Potential and total energy profiles remained constant, confirming thermal and conformational stability. Solubility profiling performed using the CamSolpH calculation server revealed that sericin 4 exhibits high solubility across a wide pH range (1-14), with a slight decrease observed around pH 10.

Overall, these results provide the first detailed computational insight into the structure, stability, and solubility properties of the sericin 4 protein. The findings not only contribute to filling the existing knowledge gap regarding this protein but also establish a reliable structural model that can serve as a foundation for future experimental studies on its biological functions, material properties, and potential biotechnological applications.

5. REFERENCES

1. R. Suryawanshi, J. Kanoujia, P. Parashar, and S. Saraf. Sericin. A versatile protein biopolymer with therapeutic significance. *Current Pharmaceutical Design* 26(42): 5414-5429 (2020).
2. G. Das, H.S. Shin, E.V.R. Campos, L.F. Fraceto, M.D.P. Rodriguez-Torres, K.C.F. Mariano, D.R. Araujo, F. Fernández-Luqueño, R. Grillo, and J.K. Patra. Sericin based nanoformulations: a comprehensive review on molecular mechanisms of interaction with organisms to biological applications. *Journal of Nanobiotechnology* 19: 30 (2021).
3. A.A. Sarymsakov, S.S. Yarmatov, and K.E. Yunusov. Extraction of Sericin from Cocoons of the Silkworm *Bombyx Mori*, Its Characteristics, and a Dietary Supplement on Its Basis to Prevent Diabetes Mellitus. *Polymer Science Series B* 66(1): 89-96 (2024).
4. M.N. Padamwar and A.P. Pawar. Silk sericin and its applications: A review. *Journal of Scientific & Industrial Research* 63(4): 323-329 (2004).
5. L. Lamboni, Y. Li, and Y. Zhang. Silk sericin-enhanced hydrogel for tissue engineering and wound healing. *Biomaterials Science* 7(11): 4567-4578 (2019).
6. Z. Wang, Y. Zhang, and Y. Yang. Sericin-based biomaterials for regenerative medicine: Current insights and future directions. *Advanced Healthcare Materials* 10(15): 2100456 (2021).
7. C.J. Park, J. Ryoo, C.S. Ki, J.W. Kim, I.S. Kim, D.G. Bae, and I.C. Um. Effect of molecular weight on the structure and mechanical properties of silk sericin gel, film, and sponge. *International Journal of Biological Macromolecules* 119: 821-832 (2018).
8. H. Yun, H. Oh, M.K. Kim, H.W. Kwak, J.Y. Lee, I.Ch. Um, S.K. Vootla, and K.H. Lee. Extraction conditions of Antheraea mylitta sericin with high yields and minimum molecular weight degradation. *International Journal of Biological Macromolecules* 52: 59-65 (2013).
9. H.Y. Kweon, J.H. Yeo, K.G. Lee, Y.W. Lee, Y.H. Park, J.H. Nahm, and C.S. Cho. Effects of poloxamer on the gelation of silk sericin. *Macromolecular Rapid Communications* 21(18): 1302-1305 (2000).
10. Y.N. Jo, B.D. Park, and I.C. Um. Effect of storage and drying temperature on the gelation behavior and structural characteristics of sericin. *International Journal of Biological Macromolecules* 81: 936-941 (2015).
11. R.I. Kunz, R.M.C. Brancalhão, L.D.F.C. Ribeiro, and M.R.M. Natali. Silkworm Sericin: Properties and Biomedical Applications. *BioMed Research International* 2016: 8175701 (2016).
12. R. Aad, I. Dragojlov, and S. Vesentini. Sericin Protein: Structure, Properties, and Applications. *Journal of Functional Biomaterials* 15(11): 322 (2024).
13. Q. Xia, Z. Zhou, C. Lu, D. Cheng, F. Dai, B. Li, P. Zhao, X. Zha, T. Cheng, C. Chai, *et al*. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306(5703): 1937-1940 (2004).
14. H. Yun, M. K. Kim, and H.W. Kwak. Structural characterization and biological activities of sericin from different silkworm races. *International Journal of Industrial Entomology* 27(1): 135-140 (2013).
15. H. Okamoto, F. Ishikawa, and Y. Suzuki. Structural analysis of sericin genes. Homologies with fibroin gene in the 5'flanking nucleotide sequences.

Journal of Biological Chemistry 257(24): 15192-15199 (1982).

16. B. Kludkiewicz, Y. Takasu, R. Fedic, T. Tamura, F. Sehnal, and M. Zurovec. Structure and expression of the silk adhesive protein Ser2 in *Bombyx mori*. *Insect Biochemistry and Molecular Biology* 39(12): 938-946 (2009).
17. K.I. Komatsu. Chemistry and structure of silk. *Jarq-Japan Agricultural Research Quarterly* 13(1): 64-72 (1979).
18. Y. Takasu, H. Yamada, and K. Tsubouchi. Isolation of three main sericin components from the cocoon of the silkworm, *Bombyx mori*. *Bioscience, Biotechnology, and Biochemistry* 66(12): 2715-2718 (2002).
19. Y. Takasu, H. Yamada, T. Tamura, H. Sezutsu, K. Mita, and K. Tsubouchi. Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*. *Insect Biochemistry and Molecular Biology* 37 (11): 1234-1240 (2007).
20. Z. Dong, K. Guo, X. Zhang, T. Zhang, Y. Zhang, S. Ma, H. Chang, M. Tang, L. An, Q. Xia, and P. Zhao. Identification of *Bombyx mori* sericin 4 protein as a new biological adhesive. *International Journal of Biological Macromolecules* 132: 1121-1130 (2019).
21. D.W. Mount (Ed.). Bioinformatics: Sequence and Genome Analysis (2nd Edition). *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, United States of America* (2004).
22. P. Chakrabarti and D. Pal. The interrelationships of side-chain and main-chain conformations in proteins. *Progress in Biophysics and Molecular Biology* 76(1-2): 1-102 (2001).
23. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583 (2021).
24. R.H. Yousif, H.A. Wahab, K. Shameli, and N.B. Khairudin. Exploring the Molecular Interactions between Neoculin and the Human Sweet Taste Receptors Through Computational Approaches. *Sains Malaysiana* 49(3): 517-525 (2020).
25. R.F. Service. The game has changed. AI triumphs at protein folding. *Science* 370(6521): 1144-1145 (2020).
26. H.A. Mesrabadi, K. Faez, and J. Pirgazi. Drug-target interaction prediction based on protein features, using wrapper feature selection. *Scientific Reports* 13: 3594 (2023).
27. K.K. Barani, M. Mohammadi, M. Ghambarian, and Z. Azizi. Fe_3O_4/ZnO@ MWCNT promoted green synthesis of biological active of new azepinooxazepine derivatives: Combination of experimental and theoretical study. *Polycyclic Aromatic Compounds* 44(1): 528-554 (2024).
28. H.A. Guvenilir and T. Doğan. How to approach machine learning-based prediction of drug/compound–target interactions. *Journal of Cheminformatics* 15: 16 (2023).
29. L.N. David, M.C. Michael, and L.L. Albert (Eds.). Polypeptides Fold Rapidly by a Stepwise Process. In: Lehninger Principles of Biochemistry (7th Edition.). *W.H. Freeman, New York, USA* (2017).
30. P. Hunter. Into the fold. Advances in technology and algorithms facilitate great strides in protein structure prediction. *EMBO Reports* 7(3): 249-252 (2006).
31. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D.A. Evans, Ch. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A.I. Cowen-Rivers, A. Cowie, M. Figurnov, F.B. Fuchs, H. Gladman, R. Jain, Y.A. Khan, C.M.R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E.D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J.M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* 630: 493-500 (2024).
32. M. Wiederstein and M.J. Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35: W407-W410 (2007).
33. M.J. Sippl. Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins* 17(4): 355-362 (1993).
34. O.V. Sobolev, P.V. Afonine, N.W. Moriarty, M.L. Hekkelman, R.P. Joosten, A. Perrakis, and P.D. Adams. A global Ramachandran score identifies protein structures with unlikely stereochemistry. *Structure* 28(11): 1249-1258 (2020).
35. E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus. Improving physical realism, stereochemistry, and

side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 77(9): 114-22 (2009).

36. L. Zhang and J. Skolnick. What should the Z-score of native protein structures be? *Protein Science* 7(5):1201-1207 (1998).
37. S. Lyskov, F.C. Chou, S.Ó. Conchúir, B.S. Der, K. Drew, D. Kuroda, J. Xu, B.D. Weitzner, P.D. Renfrew, P. Sripakdeevong, B. Borgo, J.J. Havranek, B. Kuhlman, T. Kortemme, R. Bonneau, J.J. Gray, and R. Das. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS One* 8(5): e63906 (2013).
38. G. Bayarri, P. Andrio, A. Hospital, M. Orozco, and J.L. Gelpí. BioExcel Building Blocks Workflows (BioBB-Wfs), an integrated web-based platform for biomolecular simulations. *Nucleic Acids Research* 50(W1): W99–W107 (2022).
39. M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology* 112: 531-552 (1999).
40. M. Naveed, K. Javed, T. Aziz, A. Zafar, M. Fatima, H.M. Rehman, A.A. Khan, A.S. Alamri, W.F. Alsanie, and M. Alhomrani. Innovative Approach of High-Throughput Screening in the Drug Discovery Quest for Chronic Bronchitis Treatment. *Journal of Computational Biophysics and Chemistry* 24(02): 173-187 (2025).
41. M. Naveed, I. Ali, T. Aziz, A. Saleem, Z. Rajpoot, S. Khaleel, A.A. Khan, M. Al-harbi and T.H. Albekairi. Computational and GC-MS screening of bioactive compounds from Thymus Vulgaris targeting mycolactone protein associated with Buruli ulcer. *Scientific Reports* 15(1): 131 (2025).
42. M. Oeller, R. Kang, R. Bell, H. Ausserwöger, P. Sormanni, and M. Vendruscolo. Sequence-based prediction of pH-dependent protein solubility using CamSol. *Briefings in Bioinformatics* 24(2): 1-7 (2023).