International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

CoSec: a hub of online tools for comparing secondary structure elements

Ankur Chaurasia, Jyotilipsa Mohanty, Lukkani Laxman Kumar, Ayaluru Murali

DOI: 10.1504/IJBRA.2023.10054448

Article History:

Received:	11 November 2022
Last revised:	13 November 2022
Accepted:	04 January 2023
Published online:	05 June 2023

CoSec: a hub of online tools for comparing secondary structure elements

Ankur Chaurasia, Jyotilipsa Mohanty, Lukkani Laxman Kumar and Ayaluru Murali*

Department of Bioinformatics,
School of Life Sciences,
Pondicherry University,
Pondicherry – 605 014, India
Email: ankurtchuzy@gmail.com
Email: jlipsa95@bicpu.edu.in
Email: laxman30.res@pondiuni.edu.in
Email: murali@bicpu.edu.in
*Corresponding author

Abstract: A common problem with *in-silico* protein modelling is choosing the best model out of a cluster of protein models suggested by the online servers. Besides identifying a right model based on torsion angles and potentials, a lot of researchers look at the model that retains most of its predicted secondary structures. The comparison of the secondary structure elements at residue level becomes more tedious as the size of the protein increases. So, we have developed two tools predicted secondary structure matching (PreSSM) and compare assigned secondary structure (CompASS) under one umbrella CoSec. PreSSM compares the secondary structure elements of a modelled protein from a PDB to the secondary structure predicted, while CompASS compares the secondary structures between two PDBs of the same protein (typically the models before and after simulation/docking with a ligand/mutation). Moreover, the two tools use STRIDE (with 95% consensus, with 5% divergence) to assign secondary structure confirmation to residues in the given protein's structure.

Keywords: bioinformatics; computational biology; secondary structure analysis; sequence analysis; protein structure; molecular dynamic simulation.

Reference to this paper should be made as follows: Chaurasia, A., Mohanty, J., Kumar, L.L. and Murali, A. (2023) 'CoSec: a hub of online tools for comparing secondary structure elements', *Int. J. Bioinformatics Research and Applications*, Vol. 19, No. 1, pp.56–69.

Biographical notes: Ankur Chaurasia did his graduation in Biochemistry from the University of Delhi, India and pursued a master's degree in Bioinformatics from Pondicherry University, India. He has also worked as a research trainee at the Institute of Bioinformatics, Bangalore, India. He is currently a PhD scholar at the University of Manchester, UK where he is working in the rare disease space to discover diagnosis and novel diseases genes.

Jyotilipsa Mohanty completed her graduation in Botany from Fakir Mohan Autonomous College, Odisha, India and Master's degree in Bioinformatics from Orissa University of Agriculture and Technology, Bhubaneswar, Odisha, India. She is currently pursuing PhD programme in Bioinformatics at Pondicherry University, Puducherry, India. Her research interest includes virus trafficking in plants.

Lukkani Laxman Kumar is a PhD student working at the Department of Bioinformatics, Pondicherry University, Pondicherry, India. He has graduated in Bio-technology (major) from the Nizam College (Osmania University, Hyderabad, Telangana, India) in 2015, and continued his Master's degree in Life Sciences (Bioinformatics) from Central University of Punjab, Bathinda, India in 2015; and MTech Computational Biology from Pondicherry University, Pondicherry, India in 2019. His area of research includes protein modelling, molecular dynamics simulations, transcriptomics, and computer-aided drug designing.

Ayaluru Murali is currently working as an Assistant Professor at the Department of Bioinformatics, Pondicherry University, India. He received his PhD (2000) from Sri Venkateswara University, Tirupati, India. Before taking up the position at Pondicherry University he had worked as post doc in USA. He had about 28 years of research experience. He had published about 57 papers (h-index: 28). His research interests include cryoEM, single particle analysis, and structural bioinformatics. Currently, he is actively working on finding important biological problems related to phage polymerase, plant viruses and fungi using various bioinformatics tools.

1 Introduction

Proteins are linear polymeric chains that contain combination of 20 amino acid residues. The overall shape of a protein plays an important role in determining its function. Any change in the structure of protein is reflected as a change in its activity which, sometimes, may render protein non-functional. The procedure of determining the structure of a protein to its atomic level by various methods such as X-ray crystallography, NMR and Electron microscopy is time consuming and expensive. Also, they do not always give a complete structure that can be used for further studies. The challenges faced in determining the three-dimensional structure of proteins by experimental methods have opened the doors for a new era where the structure of protein is determined by *in-silico* prediction methods. These predicted models are considered as an alternative when structure of good quality, solved by any of the structural biology tools, is not available for a protein.

A structural model can be created using template-based and template-free modelling (Abeln et al., 2017), with protein's amino acid sequence as a starting point. In template-based approach, unknown structure of a protein is modelled by homology modelling if a suitable homologous template with known structure is available. The principles of homology modelling are based on the fact that similar sequences possess similar structures and hence similar functions. Also, it is well known that the structures are more conserved than sequences during the course of evolution. If the protein lacks a template (homologous counterpart of the query protein) that covers the whole sequence, then multi-template modelling can be carried out with several templates that match segments of the query sequence. The structures of all the templates are then refined together, maintaining the initially predicted constraints as much as possible, into a complete protein structure of the query sequence (Šali and Blundell, 1993). In cases where suitable templates are not available, template-free protein structure modelling is considered as an alternative approach to predict a protein structure. Such strategies, called '*ab initio*'

strategy, use the sequence of query alone to suggest possible models of the protein. In the recent trends, AlphaFold2 – a deep learning based method, showed an excellent improvement in predicting the protein 3D structures (Nature Methods, 2022).

Each of the above two approaches predicts several models of the same protein. These models differ significantly in their secondary structure composition. Not only the size and the number of secondary structure motifs such as α-helix and the β-sheet vary, their position in the overall 3D structure also differs. The general practice carried out for selecting a single protein model out of many predicted models is to perform quality assessment of the structures. The majority of the assessment programs available online and offline are based on stereochemical, energetic and empirical quality of protein structures (Bowie et al., 1991; Colovos and Yeates, 1993; Laskowski et al., 1993; Laskowski et al., 1996; Lüthy et al., 1992; Morris et al., 1992; Pontius et al., 1996). These parameters are subject to change if a model is minimised and/or refined. Also, these assessment programs exclusively deal with the protein quality and the stability; they do not assess whether the secondary structure composition in the predicted model is appropriate and agrees to the composition that might be present in the real protein structure, a parameter that is important for a correct protein function.

Secondary structures arise due to the formation of hydrogen bonds between partially charged atoms of the polypeptide backbone. The segments of protein will either coil or fold into secondary architecture that contributes to the overall protein's shape. The dictionary of protein secondary structure (DSSP) (Kabsch and Sander, 1983) defines eight types of protein secondary structures. These include the α-helix and the β-sheet that were the first two major structural folds identified by Pauling and Corey in 1951. These motifs have turned out to be a paramount feature of protein structural assembly. Secondary structures provide a simple and intuitive description of the protein's 3D structure that can be predicted by programs. Majority of existing algorithms predict three protein states i.e., α-helix, β-strand and coils. These prediction methods have evolved from using simple single residue and segment statistics to complex evolutionary information. The first-generation prediction methods such as Chou and Fasman (1974), Lim (1974) and GORI (Garnier et al., 1978) were based on single residue statistical propensities that described the preference of a residue for a particular secondary structure state. The second-generation prediction methods such as GORIII (Gibrat et al., 1987), COMBINE and S83 combined larger database of protein structure and statistics based on segments. Typically, statistics were derived to evaluate the likeliness of a residue central to a segment of about 11–21 adjacent residues to be in a particular secondary structure state. Those algorithms were based on statistical information, physico-chemical properties, sequence patterns, multi-layered (or neural) networks, graph-theory, and nearest-neighbour algorithms.

The third-generation prediction method uses evolutionary information that are derived from alignment of multiple homologous sequences. These prediction methods employ new computational algorithms such as support vector machines, Bayesian or hidden semi-Markov network and conditional random fields for combined prediction of secondary structures. Among all the third-generation algorithms, neural-network-based models have been reported to have the highest level of accuracy. Examples of these algorithms are PSIPRED (Buchan et al., 2010), Jpred (Cole et al., 2008) and DeepCNF (Wang et al., 2016a). The accuracies of secondary structure prediction methods have reached new heights from 69.7% by PHD in 1993, 76.5% by PSIPRED in 1999, 80% by structural property prediction with integrated neural network (SPINE) in 2007, 82% by

Structural Property prediction with Integrated deep neural network 2 (SPIDER2) in 2015, to 84% by deep convolution neural field network (DeepCNF) in 2016. In the third-generation algorithms, the accuracy has inched closer towards the theoretical limit of prediction, that is about 85% to 88% (Yang et al., 2018).

Incorporating the knowledge of secondary structures can make the process of selecting a protein model out of many predicted models more reliable. A protein model that has a better agreement with predicted secondary structures will have an overall structure (and hence its function) that is closer to reality. This requires a one to one comparison of the secondary structure conformation of each residue in the given protein model with the predicted secondary structure conformation of the same residues given by the prediction methods. The conformation of the residues in the protein models can be determined using secondary structure assignment programs such as DSSP (Kabsch and Sander, 1983) and STRIDE (Frishman and Argos, 1995). However, this comparison is tedious and nearly impossible. This necessitates the requirement of a tool to perform a secondary structure comparison. Such tool can have an enormous potential to change the way a protein model is validated.

To enable the comparison of predicted secondary structure conformation of a residue with the assigned secondary structure conformation of the same residue in protein's model, we have developed an online tool named predicted secondary structure matching (PreSSM). This tool offers user a range of secondary structure prediction methods to choose from and uses STRIDE algorithm to assign secondary structure confirmations to residues in the given protein's 3D coordinates (PDB file). The result is represented as a colour coded alignment between predicted and assigned conformations of each residue along with scores to evaluate the similarity. Users can also download the results in text format for further analysis.

Once a model is selected, it can be minimised by molecular dynamics simulation and its stability can be checked at different time points of the simulation. Again, a need arises to keep a track on the changing secondary structure features in the model. For this, we have developed another online tool named compare assigned secondary structure (CompASS) that compares the secondary structure elements in the protein model at different states and gives the extent to which the states differ from each other. Also, the initial models of a protein (as seen in the homology approach and multi-template approach) can be compared to each other using this tool to observe the magnitude of dissimilarity among them. Both PreSSM and CompASS can be accessed from a common interface called compare secondary structures (CoSec) that acts as a hub with a future prospective of adding more tools for structure analysis. The tools are available at: http://cosec.bicpu.edu.in/

2 Materials and methods

2.1 Secondary structure prediction methods

PreSSM gives freedom to the users to choose from the provided prediction methods that covers all the three generations of algorithms. Although the third-generation prediction methods are the most accurate and are suggested to be used, the first-generation methods are fast and provide instant results. PreSSM is compatible with the output of four servers which include 2 third-generation secondary structure prediction servers, RaptorX

Property (Wang et al., 2016b) and PSIPRED (Buchan et al., 2013). RaptorX Property is a web server that predicts secondary structure of a protein sequence without using any templates. This server employs deep learning model deep convolutional neural fields (DeepCNF) (Wang et al., 2016a) to predict secondary structure. PSIPRED incorporates two feed-forward neural networks that analyse the output obtained from position specific iterated – BLAST (PSI-BLAST). The rationale behind this approach is to use the evolutionary knowledge of related proteins to predict the secondary structure of a new amino acid sequence.

The first-generation prediction method compatible with PreSSM include CFSSP: Chou and Fasman Secondary Structure Prediction server. This server implements Chou-Fasman algorithm (Chou and Fasman, 1974) that analyses the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on solved protein structures. These frequencies are used to derive a set of probability parameters that are used to predict the probability that a given sequence of amino acids would form a certain secondary structure fold. Another server that can be used to obtain secondary structure prediction is NSPA server (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html). It gives users an option to choose one, or combination of many secondary structure prediction methods listed therein and then the consensus prediction program generates a prediction. In the consensus, the most abundant predicted conformational state is reported for each residue. Secondary structure prediction methods available at NSPA server are: DPM, DSC, GORI, GORIII, GORIV, HNN, MLRC, PHD, PREDATOR, SIMPA96, SOPM, SOPMA.

2.2 Secondary structure assignment method

Both PreSSM and CompASS use STRIDE which is a software tool for secondary structure assignment, given the atomic-resolution coordinates of the protein. This method applies a knowledge-based algorithm that makes combined use of hydrogen bond energy and statistically derived backbone torsional angle information (Frishman and Argos, 1995). The STRIDE web server [24] provides a channel to both the tools to access STRIDE algorithm – The assigned conformations are: Alpha helix (H), 3–10 helix (G), PI-helix (I), Extended conformation (E), Isolated bridge (B or b), Turn (T) and Coil (C). Although, DSSP algorithm is considered as the standard method for assigning secondary structure to the residues of a protein, STRIDE is used here. This is because the STRIDE considers both hydrogen bonding patterns and backbone geometry while DSSP identifies only the intra-backbone hydrogen bonds using a purely electrostatic definition.

2.3 Principle of PreSSM

A simplified flow-chart of the underlying process is described in Figure 1. The first step to perform a comparison between predicted secondary structure and assigned secondary structure is to get the prediction from the listed servers (see Figure 2(A)).

The steps to be followed for doing this are explained in depth in readme file of the tool. Once the prediction is obtained, it can be simply given as the first input by copying it and pasting it in the specified textbox (see Figure 2(B)). The second input that the users have to give is the atomic coordinate file of the 3D model. The file can be opened with any text editor and the content has to be pasted in the second text area (see Figure 2(B)).

Figure 1 Flow chart showing the working principle of PreSSM (see online version for colours)

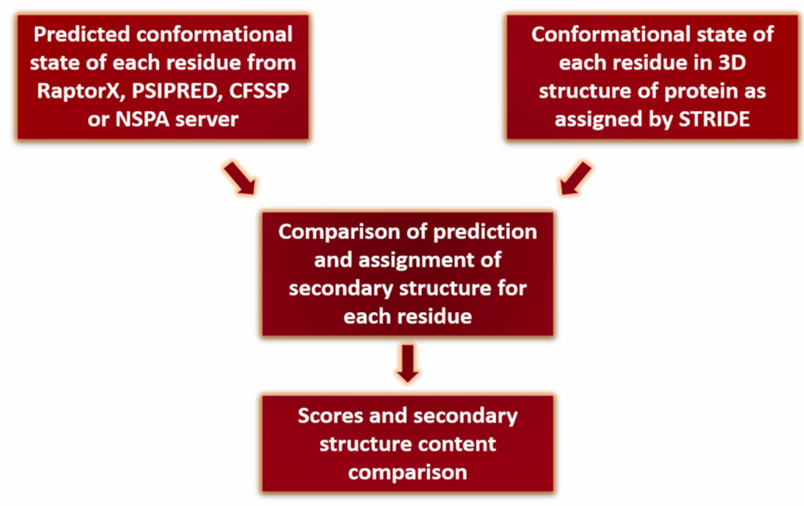
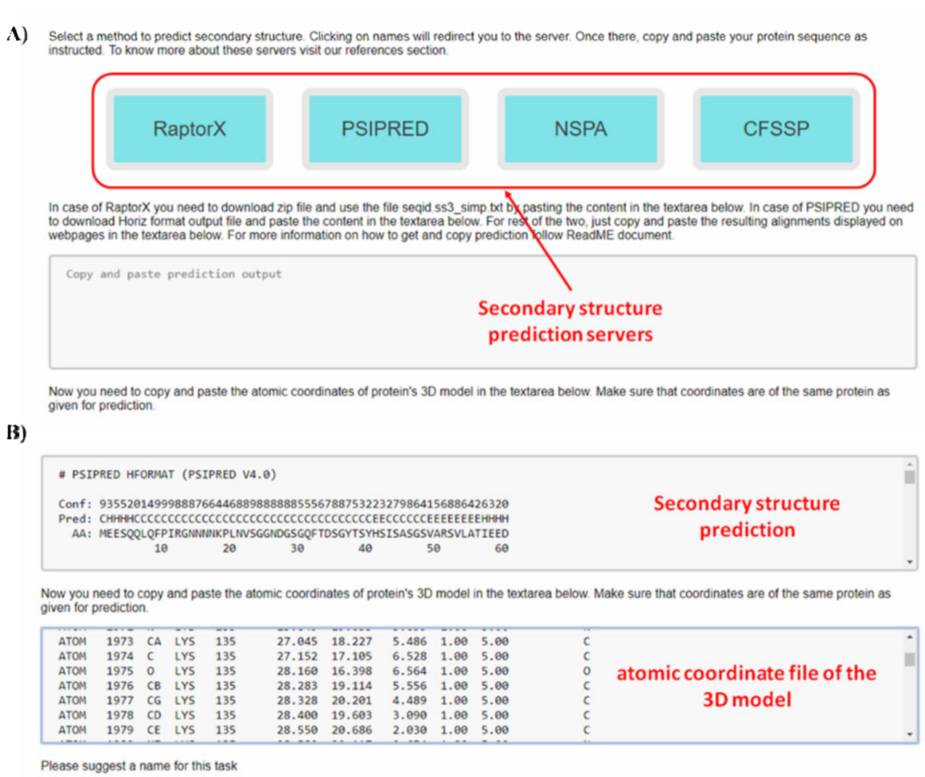


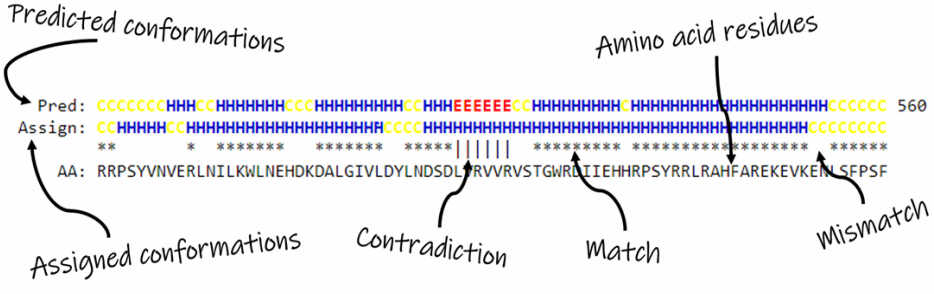
Figure 2 PreSSM webpage: (A) the secondary structure prediction servers available at PreSSM and (B) the sample input fields for prediction result and 3D coordinate file (see online version for colours)



2.3.1 Outcome of PreSSM

Upon submission, PreSSM displays an alignment that shows the secondary structure conformation of residues as suggested by prediction and the assigned conformation of the residues in 3D model along with scores (see Figure 3). The resulting alignment displays the following information:

Figure 3 Sample representation of PreSSM output alignment. The first two lines show the secondary structure conformation from prediction and assignment for each amino acid (shown in 4th line). The presence of '*' (in third line) confirms a match between predicted and assigned secondary structure conformation. The residues that have contradicting secondary structure conformations are represented by a vertical line '|' (see text for details) (see online version for colours)



2.3.2 Match score

Match score is the percentage of residues that have the same secondary structure state given in prediction as well as in 3D model secondary structure assignment. A match is represented by "*" in alignment. A model with good agreement between prediction and assignment gives a higher match score.

2.3.3 Contradiction score

Contradiction score is the percentage of residues that have altered secondary structure states i.e., helix conformational state in prediction becoming a sheet conformational state in 3D model secondary structure assignment or vice versa. A contradiction is denoted by " | " in alignment. For a good agreement between prediction and assignment, contradiction score should be low. However, the residues forming a coil changing to a helix or sheet conformation is often observed and hence will not contribute to contradiction.

2.3.4 Secondary structure content

Helix, sheet, and coil content is the percentage of residues in respective conformational states in prediction as well as in 3D model secondary structure assignment. A difference between these percentage is also given to observe deviation between prediction and assignment.

2.4 Principle of CompASS

CompASS is a tool that matches the secondary structure elements of a protein, with the secondary structure elements of the same protein in a different state by taking the atomic coordinates of both the states as inputs (see Figure 4). Users can submit the two coordinate files (PDB files) of the same protein along with name tags to make their identification easier in resulting alignment (see Figure 5). The assignment of secondary structure in this tool is based on STRIDE.

Figure 4 Flow chart showing the working principle of CompASS (see online version for colours)

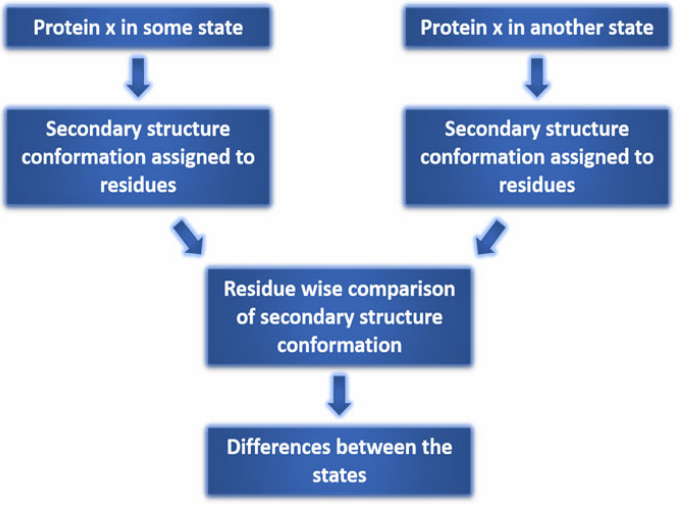
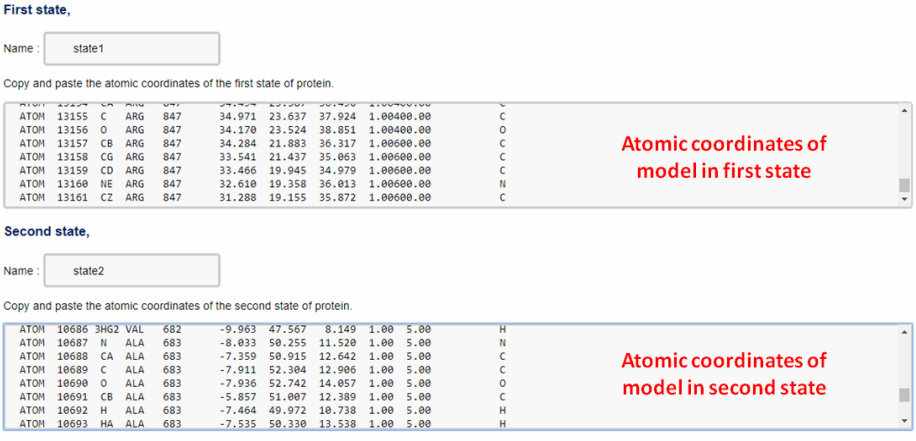


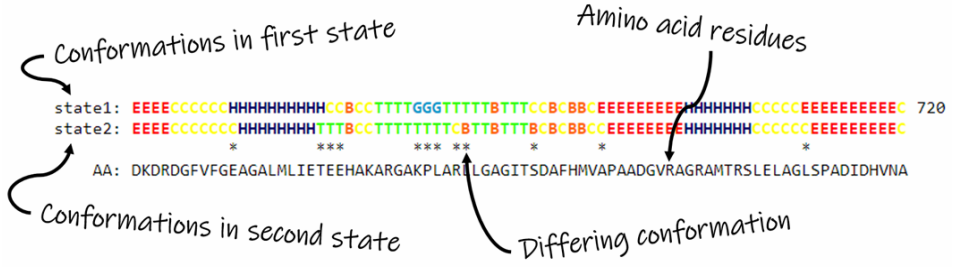
Figure 5 A sample data for running CompASS tool. The user is expected to paste the pdbs of two states (in text form) in respective boxes (see online version for colours)



2.4.1 Outcome of CompASS

Once structural coordinates of the two states of a protein are submitted, the tool uses STRIDE server to assign secondary structure conformation to the residues in both the models. This assignment is aligned to each other along with the protein sequence (that is identical in both the models) and those residues are marked that have a difference in assigned conformation (see Figure 6). The alignment result is colour coded for all the seven types of assigned conformations to make it visually perceptible.

Figure 6 Sample representation of CompASS output alignment. The first two lines show the secondary structure conformation assigned to each amino acid residue (shown in 4th line) of the submitted models. The presence of '*' (in third line) indicates a mismatch (see online version for colours)



2.4.2 Differing score

Differing Score is the percentage of residues that have different secondary structure conformation in the given two states of protein. Residues that differ in conformation are represented by "*" in alignment.

Along with the differing scores, the content percentage of individual secondary structure conformation (Alpha helix, 3-10 helix, PI-helix, Extended conformation, Isolated bridge, Turn and Coil) in both the states and the difference between them is also given.

3 Results

3.1 PreSSM

To demonstrate the practical application of PreSSM a protein with unknown experimental structure was modelled using *ab initio* method. Sequence of Eggshell organising factor 1 (EOF1, AAEL012336) was extracted from UniProt and was submitted to Robetta server. EOF1 has an essential role in eggshell melanisation and embryonic development in *Aedes aegypti* mosquitoes (Isoe et al., 2019). Five models with significant differences with regard to secondary structures were generated by the Robetta server. These models were subjected to assessment on PreSSM web interface. The secondary structure prediction method selected for this case was PSIPRED. The result of comparison is presented in Table 1.

Table 1 Comparison of EOF1 protein secondary structure (PSIPRED) with the modelled Robetta structures using PreSSM

PreSSM	*Match score*	*Contradiction score*
Model 1	67.45%	2.36%
Model 2	69.46%	2.59%
Model 3	64.98%	2.36%
Model 4	67.10%	2.48%
Model 5	67.10%	2.59%

From the above scores, it is evident that model 2 has the highest match score and contradiction score. As the contradiction score of this model differs roughly about 0.2 percentage from the lowest contradiction score, it is less significant in this case. So, taking the match score into account, model 2 has the best agreement with the predicted secondary structure and hence will be closer to the protein structure present in nature as compared to other models.

3.2 CompASS

The application of CompASS is demonstrated here with the help of the same five protein models generated for EOF1 using ROBETTA server. All the models were compared with each other using this online tool. The differences in secondary structure elements obtained by CompASS are represented in Table 2.

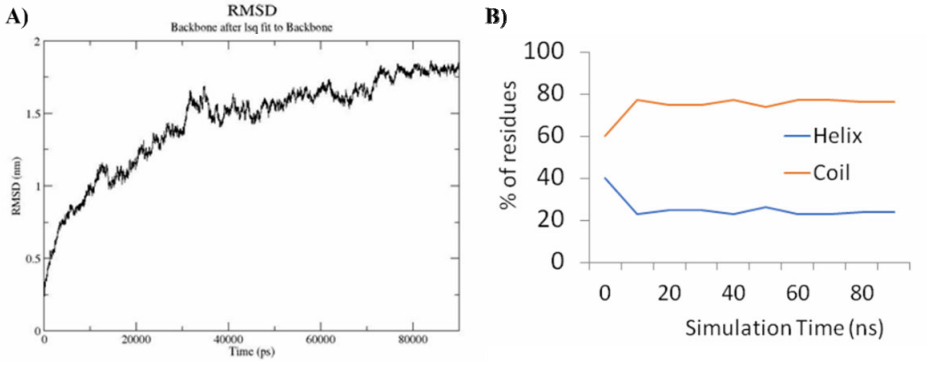
Table 2 Comparison of EOF1 protein models with each other for difference in secondary structure elements and the differing scores obtained from CompASS were represented in tabular form

CompASS	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Model 1	0	46.23%	46.23%	45.87%	44.34%
Model 2	46.23%	0	46.46%	50.71%	45.87%
Model 3	46.23%	46.46%	0	52.00%	44.69%
Model 4	45.87%	50.71%	52.00%	0	49.76%
Model 5	44.34%	45.87%	44.69%	49.76%	0

From Table 2, it is apparent that the models given by ROBETTA are very different from each other. This difference makes it difficult to choose the correct model. One way to identify a correct model is to compare its secondary structure elements with prediction using PreSSM. To validate the CompASS tool, molecular dynamics simulations were carried out on EOF1 using GROMACS software (Abraham et al., 2015). The simulation was carried out for duration of 90 ns which resulted in a stable structure, as evident from its RMSD profile (Figure 7(A)). The structures were extracted after every 10 ns during the 90 ns simulation period and the secondary structure elements were analysed using CompASS (data not shown). The number of residues constituting the helices and coils were plotted against the cumulative simulation time (Figure 7(B)). It can be noticed that the secondary structure composition stabilises after a simulation time of 60 ns which correlates well with the RMSD profile. As is evident, the structure with stabilised RMSD

indicates a stable conformation and hence the secondary structure composition should be invariant during this period. The fact that this stability in secondary structure composition is observed in the CompASS analysis demonstrates the validity of the CompASS tool.

Figure 7 MD simulations of EOF1 protein: (A) The RMSD profile of EOF1 run for 90 ns simulation period. It can be noticed that the structure was stabilised after a simulation time of 60 ns. (B) The secondary structure analysis was performed with CompASS for the structures extracted at every 10 ns interval of the simulation. A stability in the secondary structure element composition during 60–90 ns period (where the RMSD also showed a stability) validates the analysis of CompASS (see online version for colours)



4 Discussion

Protein structure prediction has gained importance in situations where the experimentally determined structures are not known. Prediction tells the possibilities, i.e., it gives the probable models of the protein by analysing the sequence. This results in a cluster of models that most often differ from each other. Choosing the correct model out of many possibilities often becomes challenging with the proteins with long sequence. A possible and logical approach is to choose the protein model that has the true secondary structure composition. This can be done by comparing secondary structure of all the models with the secondary structure prediction that are accurate up to 84% (Yang et al., 2018). This has been made possible with PreSSM. The scores given by this tool that evaluate the similarity between predicted and assigned secondary structure conformations include match score and contradiction score. A match score is the percentage of residues that have identical confirmations in both the cases. A contradiction score is the percentage of residues that have a helical confirmation in prediction but a sheet confirmation in assignment and vice versa. Both the scores can be used together to validate a model. This has been demonstrated by comparing predicted models of a protein to a secondary structure prediction. The tool offers wide options for secondary structure prediction; we recommend users to go for RaptorX server for secondary structure prediction as it has been proven to be highly accurate.

The difference between the initial models can be observed by comparing the secondary structure elements present in them. This can be done by submitting the coordinates of the models to CompASS. The difference in structural elements is described by the differing score that is the percentage of residues that have different

secondary structure conformation in the given two models of the same protein. The differences between the initial predicted models of a protein were tested using this tool. From the Table 2, larger differences can be observed when compared with other predicted models. This large deviation is primarily due to the absence of template structure for the given input sequence. Also, with the help of molecular dynamics simulations of EOF1, the stability of the secondary structure composition was monitored using CompASS and a stable composition was observed during the simulation where the RMSD profile exhibited the stability.

5 Conclusions

The current pitfall in the selection procedure of a protein model from many predicted ones can be solved by comparing the secondary structure content of these models with the predicted secondary structure. This will lead to selection of a protein model with the secondary structure composition that may relate to the real protein in nature. To perform this simple comparison, two tools have been developed and made available at a central web interface called CoSec.

The first tool, named as PreSSM, compares the predicted secondary structure conformation of each amino acid residue in a protein with the conformational state of the same residues in protein's given 3D structure. A common use of this tool can be a comparison of secondary structure in different structural models of same protein with the prediction to choose the best model among them.

The second tool, CompASS, compares conformational state of the residues in given two models of the same protein. This tool uses STRIDE algorithm for assigning the secondary structure conformation to the residues in both models for comparison. The result of comparison is aligned and a quantitative score that helps in evaluation of the difference will be provided. These two tools will be useful to scientists working in molecular biology and bioinformatics.

5.1 Usage of CoSec

Both the tools are freely available under the common web interface 'CoSec' at the link http://cosec.bicpu.edu.in/. In order to use the tools, users have to register themselves first. Both the tools are made simple to use by providing necessary documents with step by step instructions. The 'about' section of both the tools give an overview and the interpretation of the outputs. The result of comparison is displayed as a colour coded alignment to make it visually perceptible. Results of both the tools can also be downloaded in text format for future reference. The downloaded result will contain both the alignment as well as quantitative scores for comparison.

Acknowledgements

The authors thank the Department of Bioinformatics, Pondicherry University, for providing the computational facility to carry out the work.

Author contributions

AC: Conceptualisation, Methodology, Software, Investigation, Writing – Original Draft.

JM: Investigation, Validation, Editing the manuscript.

LLK: Investigation, Validation, Editing the manuscript.

AM: Conceptualisation, Writing – Review and Editing, Resources, Supervision, Project administration.

Availability and implementation

The web server of CoSec is freely available at http://cosec.bicpu.edu.in/

References

Abeln, S., Heringa, J. and Feenstra, K.A. (2017) 'Strategies for protein structure model generation', *Structural Bioinformatics*, pp.7–17, https://doi.org/10.48550/arXiv.1712.00425

Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B. and Lindahl, E. (2015) 'GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers', *SoftwareX*, Vols. Vol. 1, No. 2, pp.19–25, https://doi.org/10.1016/j.softx.2015.06.001

Bowie, J., Luthy, R. and Eisenberg, D. (1991) 'A method to identify protein sequences that fold into a known three-dimensional structure', *Science*, Vol. 253, No. 5016, pp.164–170, https://doi.org/10.1126/science.1853201

Buchan, D.W.A., Ward, S.M., Lobley, A.E., Nugent, T.C.O., Bryson, K. and Jones, D.T. (2010) 'Protein annotation and modelling servers at university college london', *Nucleic Acids Research*, Vol. 38, Web Server, pp.W563–W568, https://doi.org/10.1093/nar/gkq427

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K. and Jones, D.T. (2013) 'Scalable web services for the PSIPRED protein analysis workbench', *Nucleic Acids Research*, Vol. 41, No. W1, pp.W349–W357, https://doi.org/10.1093/nar/gkt381

Chou, P.Y. and Fasman, G.D. (1974) 'Prediction of protein conformation', *Biochemistry*, Vol. 13, No. 2, pp.222–245, https://doi.org/10.1021/bi00699a002

Cole, C., Barber, J.D. and Barton, G.J. (2008) 'The Jpred 3 secondary structure prediction server', *Nucleic Acids Research*, Vol. 36, Web Server, pp.W197–W201, https://doi.org/10.1093/nar/gkn238

Colovos, C. and Yeates, T.O. (1993) 'Verification of protein structures: patterns of nonbonded atomic interactions', *Protein Science*, Vol. 2, No. 9, pp.1511–1519, https://doi.org/10.1002/pro.5560020916

Frishman, D. and Argos, P. (1995) 'Knowledge-based protein secondary structure assignment', *Proteins: Structure, Function, and Genetics*, Vol. 23, No. 4, pp.566–579, https://doi.org/10.1002/prot.340230412

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) 'Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins', *Journal of Molecular Biology*, Vol. 120, No. 1, pp.97–120, https://doi.org/10.1016/0022-2836(78)90297-8

Gibrat, J-F., Garnier, J. and Robson, B. (1987) 'Further developments of protein secondary structure prediction using information theory', *Journal of Molecular Biology*, Vol. 198, No. 3, pp.425–443, https://doi.org/10.1016/0022-2836(87)90292-0

Isoe, J., Koch, L.E., Isoe, Y.E., Rascón, A.A., Brown, H.E., Massani, B.B. and Miesfeld, R.L. (2019) 'Identification and characterization of a mosquito-specific eggshell organizing factor in aedes aegypti mosquitoes', *PLoS Biology*, Vol. 17, No. 1, pp.1–23, https://doi.org/10.1371/journal.pbio.3000068

Kabsch, W. and Sander, C. (1983) 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers*, Vol. 22, No. 12, pp.2577–2637, https://doi.org/10.1002/bip.360221211

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) 'PROCHECK: a program to check the stereochemical quality of protein structures', *Journal of Applied Crystallography*, Vol. 26, No. 2, pp.283–291, https://doi.org/10.1107/S0021889892009944

Laskowski, Roman, A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) 'AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR', *Journal of Biomolecular NMR*, Vol. 8, No. 4, pp.477–486, https://doi.org/10.1007/BF00228148

Lim, V.I. (1974) 'Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure', *Journal of Molecular Biology*, Vol. 88, No. 4, pp.857–872, https://doi.org/10.1016/0022-2836(74)90404-5

Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992) 'Assessment of protein models with three-dimensional profiles', *Nature*, Vol. 356, No. 6364, pp.83–85, https://doi.org/10.1038/356083a0

Nature Methods (2022) 'Method of the Year 2021: Protein Structure Prediction', *Nature Methods*, Vol. 19, No. 1, p.1, https://doi.org/10.1038/s41592-021-01380-4

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) 'Stereochemical quality of protein structure coordinates', *Proteins: Structure, Function, and Bioinformatics*, Vol. 12, No. 4, pp.345–364, https://doi.org/10.1002/prot.340120407

Pontius, J., Richelle, J. and Wodak, S.J. (1996) 'Deviations from standard atomic volumes as a quality measure for protein crystal structures', *Journal of Molecular Biology*, Vol. 264, No. 1, pp.121–136, https://doi.org/10.1006/jmbi.1996.0628

Šali, A. and Blundell, T.L. (1993) 'Comparative protein modelling by satisfaction of spatial restraints', *Journal of Molecular Biology*, Vol. 234, No. 3, pp.779–815, https://doi.org/10.1006/jmbi.1993.1626

Wang, S., Peng, J., Ma, J. and Xu, J. (2016a) 'Protein secondary structure prediction using deep convolutional neural fields', *Scientific Reports*, Vol. 6, January, pp.1–11, https://doi.org/10.1038/srep.18962

Wang, S., Li, W., Liu, S. and Xu, J. (2016b) 'RaptorX-property: a web server for protein structure property prediction', *Nucleic Acids Research*, Vol. 44, No. W1, pp.W430–W435, https://doi.org/10.1093/nar/gkw306

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K. and Zhou, Y. (2018) 'Sixty-five years of the long march in protein secondary structure prediction : the final stretch ?', *Briefings in Bioinformatics*, Vol. 19, pp.482–494, December 2016, https://doi.org/10.1093/bib/bbw129