Applied Chemical Engineering (2025) Volume 8 Issue 1 ISSN: 2578-2010 (O)
doi: 10.59429/ace.v8i1.5588

ORIGINAL RESEARCH ARTICLE

Deep learning approach for secondary structure prediction of Pseudomonas aeruginosa –Quorum sensing repressor

Saravanan K[1,*], Sivakumar S[2], Marimuthu T[3], Palanisamy P.N[4], Sangeetha P[5], Sangeetha B[6]

[1] *Department of Physics, AVS Engineering College, Salem, Tamilnadu 636003, India*

[2] *Department of Physics, Government Arts College (Autonomous), Salem, Tamilnadu 636007, India*

[3] *Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education (Deemed to be University), Krishnankoil, Tamilnadu 626126, India*

[4] *Department of Electronics and Communication Engineering, Mahendra College of Engineering, Salem, Tamilnadu 636106, India*

[5] *Department of Physics, Sona College of Technology, Salem, Tamilnadu 636005, India*

[6] *Department of Electrical and Electronics Engineering, AVS Engineering College, Salem, Tamilnadu 636003, India*

***Corresponding author:** Saravanan.K, saravanan.avs20@gmail.com

ABSTRACT

Accurate prediction of Protein Secondary Structure (PSS) plays a crucial role in understanding the functional mechanisms of proteins. This study focuses on predicting the secondary structure of the quorum-sensing control repressor protein (QscR) using a UNet based deep learning model. The UNet architecture, known for its exceptional performance is adapted to predict structural features of proteins by learning from sequence based data. The proposed model was trained and validated using benchmark protein datasets to ensure generalizability and accuracy. Comparative analysis with traditional approaches demonstrated that the UNet model achieved superior performance in terms of prediction accuracy and computational efficiency. The findings suggest that the UNet model is a robust tool for SS prediction and can provide deeper insights into quorum-sensing mechanisms, aiding in the design of novel antibacterial strategies.

Keywords: *pseudomonas aeruginosa*; quorum-sensing repressor; secondary structure prediction; UNet; bacterial proteins

ARTICLE INFO

Received: 25 December 2024
Accepted: 17 April 2025
Available online: 09 May 2025

1. Introduction

Proteins are indispensable to living organisms due to their diverse functions, including catalyzing essential reactions in cellular metabolism, participating in DNA replication, and producing antibodies for the immune system. These vital biomolecules exhibit a hierarchical structure comprising four levels: primary, secondary, tertiary, and quaternary. Each structural level plays a crucial role in determining the protein's overall function and significance within living systems.

Understanding protein structure is a key to elucidating its function. To comprehend the molecular-level roles of proteins, it is essential to determine their secondary structure (SS)[1]. However, accurately and reliably predicting protein structures from amino acid sequences remains one of the most challenging tasks in computational biology. This complexity hampers the analysis of protein functions and their applications in drug design. Therefore, predicting protein secondary structure (PSS) is a critical step toward tertiary structure prediction, as

it provides valuable insights into protein activity, relationships, and functions.

In bioinformatics, protein structures are commonly determined using various experimental techniques, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and computational methods. However, conventional approaches often rely on simplistic models and assumptions that fail to capture the complexity of protein folding. These methods typically analyze individual sequences in isolation, without incorporating evolutionary information available through multiple sequence alignments or Position-Specific Scoring Matrices (PSSMs).

To address these limitations, machine learning and deep learning techniques have been introduced[2-5]. These advanced approaches leverage large datasets, evolutionary insights, and sophisticated models to deliver more accurate and robust predictions, effectively overcoming many of the shortcomings of traditional methods.

A two-dimensional fusion deep neural network model, DstruCCN, which integrates CNNs with a supervised Transformer-based protein language model for single-sequence protein structure prediction. The training features extracted from both models are fused to predict the protein's Transformer binding site matrix, followed by three-dimensional structure reconstruction through energy minimization techniques. This hybrid approach leverages the local feature extraction capabilities of CNNs and the contextual understanding of Transformers, aiming to improve prediction accuracy from individual sequences. However, despite its advantages, the model also has certain limitations. The reliance on single-sequence input can lead to reduced performance compared to methods that utilize multiple sequence alignments (MSAs) or evolutionary profiles[6].

A hybrid architecture combining CNNs and LSTM networks, referred to as EN-CSLR has been formulated. Feature maps extracted from the second convolutional layer are passed through a softmax classifier to generate the first set of probability outputs. In parallel, the LSTM model comprises a sequence processing layer and a final layer, from which features are extracted and fed into a Random Forest classifier to produce the second probability output. These two probabilistic outputs are then weighted and integrated to form the final prediction. The model's effectiveness was validated using cross-validation experiments on the 25pdb dataset, achieving an accuracy of 80.18%, which surpasses the performance of individual CNN or LSTM models. Despite its strengths, integration of multiple models increases the system's complexity, leading to higher computational costs and potentially longer training times[7].

A DL framework named Cascaded Feature Learning Model (CFLM) for PSS prediction. The proposed model utilizes a multi-stage transfer learning approach built upon the Residual Dense Network (RDN), enabling progressive refinement of learned features across different training stages. The effectiveness of CFLM is validated on the CASP 13 and CASP 14 benchmark datasets. Furthermore, comparative analysis shows that CFLM outperforms several recent PSSP methods, highlighting its competitive edge in prediction accuracy. But, the multi-stage training and transfer learning process increases training time and resource demands, making the model computationally intensive. Additionally, while transfer learning enhances generalization, it may also introduce overfitting risks[8].

A Neural network (NN) based prediction has been developed[9,10]. A concept of adversarial learning by proposing a Conditional Generative Adversarial Network (CGAN)-based model has been formulated[11]. The architecture incorporates a specially designed multiscale convolution module and an Improved Channel Attention (ICA) module within the generator, enhancing its ability to extract and focus on intricate protein features. The proposed CGAN-PSSP method, driven by the multiscale and attention-based enhancements, demonstrates the potential of adversarial learning in capturing subtle and high-level features of protein sequences. Experimental results validated the feasibility and effectiveness of the CGAN-PSSP model, representing its strong feature learning capability and suggesting that it is a promising direction for future research in PSSP. But, training GANs, particularly in the context of bioinformatics, can be unstable and sensitive to hyperparameter tuning, often requiring careful balancing between the generator and discriminator.

The influence of amino acid properties and SSPs in predicting secondary structure was formulated. This model first utilizes D-Conv and a SENet for capturing local patterns and enhancing relevant channels. It then integrates recurrent NN variants along with a Transformer module, to extract global bidirectional dependencies and refine feature representation. However, the use of multiple DL components adds to the computational complexity and may pose challenges in terms of scalability and resource requirements[12].

DL architectures such as CNN, RNN, Inception Networks and Graph Neural Networks (GNNs) have been widely adopted in PSSP. Additionally, techniques originally developed for natural language processing (NLP) and computer vision have also been successfully applied to capture both local and global dependencies in protein sequences[13,14].

These advancements in PSSP have leveraged DL architectures to improve prediction accuracy. While hybrid models can effectively capture both local and global dependencies, they often suffer from increased computational complexity and training instability. In this context, U-Net emerges as a promising alternative due to its encoder-decoder structure and skip connections, which allow it to efficiently retain both fine-grained and high-level features. It offers a balanced trade-off between performance and resource efficiency, making it well-suited for PSSP tasks. So this work opted UNet for PSS prediction.

2. Methodology

Figure 1 illustrates a workflow for PSSP using a UNet model. It begins with collecting protein data from reliable sources, followed by dataset preparation and extraction of essential features like PSSM, SSPs, and physicochemical properties. These features are input into the UNet model, which captures both local and global dependencies through its encoder-decoder structure with skip connections. The model then predicts the secondary structure of proteins, aiding in structural and functional understanding.

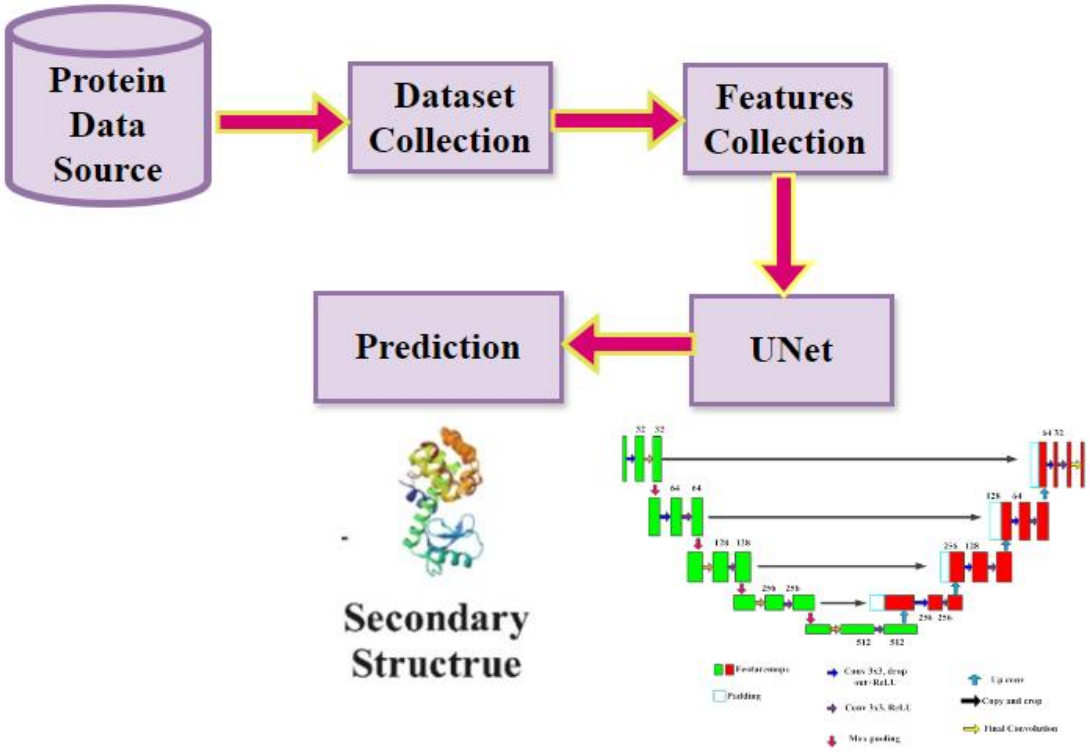

Figure 1. Methodology of the proposed topology.

3. Dataset

The Pseudomonas Genome Database (https://www.pseudomonas.com) is a specialized online resource designed for comprehensive genomic analysis of *Pseudomonas* species, especially *Pseudomonas aeruginosa*. It is widely used by microbiologists and bioinformaticians for studying genes, proteins, regulatory pathways, and virulence factors. Pseudomonas aeruginosa strains typically have between 5,500 and 6,000 open reading frames (ORFs), which correspond to protein-coding genes.

After selecting the dataset, it is essential to split it into training, validation, and test sets to effectively evaluate the model's performance. This data splitting process ensures that the model generalizes well to unseen data and helps prevent over fitting.

4. Data pre-processing

The preprocessing stage was crucial in preparing the protein sequence data for model training and evaluation. Initially, protein sequences were retrieved from the Database. Sequences with incomplete annotations, ambiguous residues, or significant redundancy were filtered out to ensure data quality and diversity.

Secondary structure annotation was performed using the DSSP (Define Secondary Structure of Proteins) tool. DSSP assigns structural states to each residue based on hydrogen bond patterns and geometric criteria derived from protein 3D structures. The original eight DSSP structural states were simplified into a standard three-state classification: H (Helix), E (Strand) and C (Coil).

Each protein sequence was processed using a sliding window technique to capture the local sequence context for each residue. The central residue within each window was labeled with its corresponding secondary structure. This annotated dataset was then split into training and testing sets, maintaining a consistent structural class distribution.

In this study, 70% of the dataset was used for training, while the remaining 10% and 20% each was allocated for validation and testing purposes.

The structural assignment of the training data and testing data set are represented in **Figure 2**, and **Figure 3**.

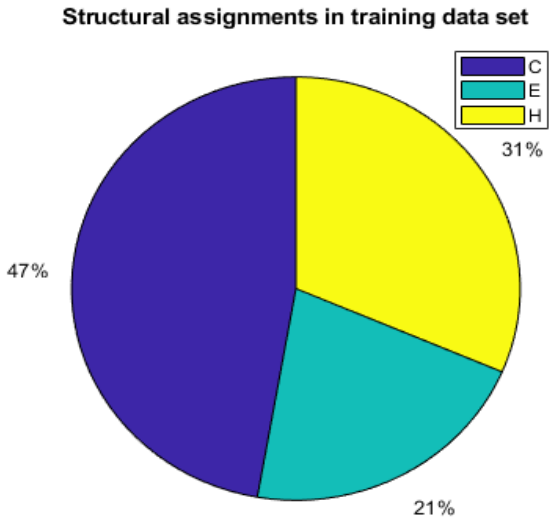

Figure 2. Structural assignments in training data set.

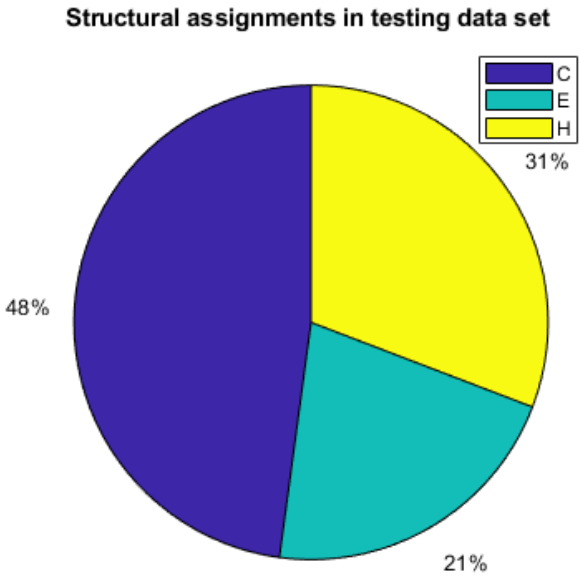

Figure 3. Structural assignments in testing data set.

From **Figures 2** and **3**, it is observed that the structural assignments in the training dataset consist of 47% coil, 31% strand, and 21% helix. In comparison, the testing dataset contains 48% coil, 31% strand, and 21% helix.

After splitting the dataset, each AA in the training and testing sequences was transformed into a numerical representation suitable for input into the DL model. One-hot encoding technique was incorporated to convert protein sequences from text to numerical data using this topology and each AA is represented by a binary vector whose length is about 20. To incorporate evolutionary information, Position-Specific Scoring Matrices (PSSMs) were generated using PSI-BLAST. Each residue was represented by a 20-dimensional vector capturing the likelihood of amino acid substitutions, based on multiple sequence alignments. The one-hot encoded vector and the corresponding PSSM values were concatenated to form a single 40-dimensional feature vector for each AA. Each window was then represented as a 17×40 matrix, capturing both sequence and evolutionary features of the surrounding AAs. These matrices were compiled into a 3D tensor of shape. This tensor structure served as the input for the DL model, which expects two-dimensional feature maps[12-15]. In this work, UNet is incorporated as DL topology, to find the SS of protein.

5. UNet architecture

Among various architectures, U-Net has emerged as a powerful model for sequence-based tasks due to its ability to learn both global context and local details[15,16]. Originally developed for biomedical image segmentation, U-Net employs a symmetric encoder–decoder structure with skip connections that allow precise localization and contextual understanding. The architecture of 2D UNet is depicted in **Figure 4**.

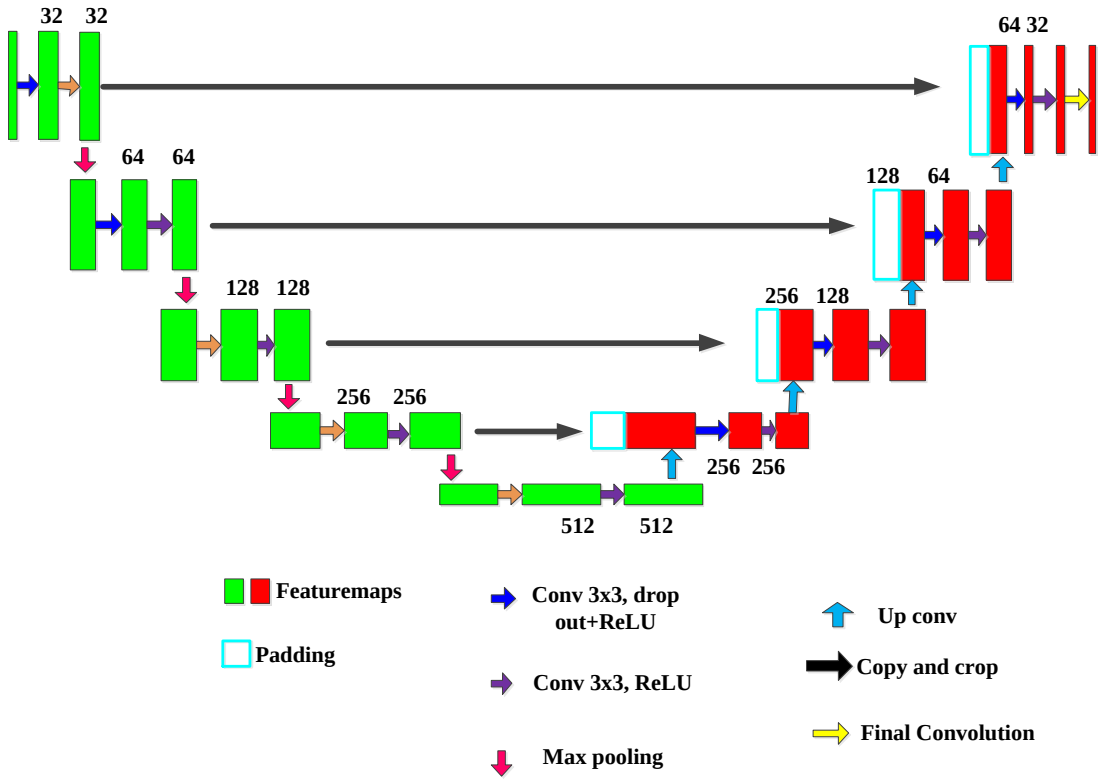

Figure 4. Architecture of 2D UNet.

To effectively capture complex spatial and contextual dependencies, the model accepts a 17×40 input matrix, integrating sequence, evolutionary and physicochemical features. Each sample has a sequence length of 128 and a feature dimension of 20, resulting in a consistent input shape suitable for UNet processing. The UNet framework is augmented with up to 512 convolutional filters in the deepest layers to enhance the model's capacity for feature representation. The network begins with a base of 32 filters and employs a consistent stride of 1 to preserve spatial resolution. A learning rate of 0.001 was selected after experimentation, offering stable convergence with the Adam optimizer. A batch size of 1 was chosen due to memory constraints and the sequential nature of protein data. Class balanced categorical cross-entropy loss was employed to address class imbalance, ensuring fair learning across all categories. ReLU activation was used for its computational efficiency and ability to mitigate vanishing gradients. Min-Max scaling was applied to normalize feature ranges, aiding in faster convergence. A stride of 1 was maintained throughout to preserve spatial detail. Feature fusion is achieved using 1×1 convolution layers, which help reduce dimensionality and facilitate interaction across channels. Dropout-augmented convolutional blocks are incorporated to reduce overfitting by randomly deactivating neurons during training.Early stopping was used to improve generalization. Thus, the training parameters utilized for segmentation is portrayed in **Table 1**.

Table 1. Training parameters for segmentation model.

Parameters	**U-Net**
Loss	Categorical crossentropy
Batch size	1
Stride	1
Early stopping	Y
Learning rate	0.001
Number of base filters	32
Scaling	Max-Min scaling

Parameters	U-Net
Optimizer	Adam optimizer
Activation Function	ReLu
sequence_length	128
feature_dimension	20
num_classes	3

Table 1. (*Continued*)

6. Results and discussion

The proposed UNet model was trained on a dataset of protein sequences with annotated secondary structures and validated using the *Pseudomonas aeruginosa* quorum-sensing repressor protein. **Table 2** depicts the comparative distribution of secondary structure elements alpha helices, beta sheets, and coils predicted by different methods including DSSP[17], STRIDE[18], and the proposed UNet-based model.

Table 2. Secondary structure distribution between methods.

Method	Alpha (%)	Beta sheet(%)	Coil (%)
DSSP	54	13	-
STRIDE	54	11	-
UNet	57	14	29

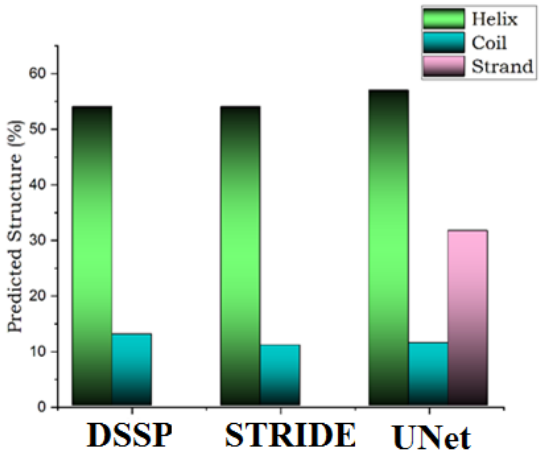

Figure 5. Secondary structure distribution between methods.

From the **Table 2**, it is observed that the UNet based model demonstrates strong alignment with DSSP and STRIDE in predicting alpha helices and beta sheets, with minor variations. The predicted alpha-helix content from UNet is slightly higher (57%) compared to DSSP and STRIDE (54%). Similarly, beta sheet prediction is close: 14% (UNet) vs. 13% (DSSP) and 11% (STRIDE). The coil percentage (29%) is only available from UNet due to limitations in the tabulated outputs of DSSP and STRIDE. In practice, DSSP and STRIDE do account for coil/loop regions but often focus reporting on structured regions (H/E). The presence of a non-zero coil prediction from UNet suggests the model's ability to identify flexible or unstructured regions, which is important for biological relevance particularly in proteins like quorum-sensing repressors that may contain flexible DNA-binding.

Thus, the performance metrics of the proposed system is tabulated in **Table 3**.

Table 3. Performance measures.

Structure	Precision	Recall	F1-score
H	0.877	0.926	0.901
E	0.857	0.923	0.889
C	0.931	0.818	0.871

The results demonstrate that the UNet model exhibits high precision and recall across all structure types, with the best F1-score of 0.901 achieved for helices, which constitute a dominant portion of the protein. The strand predictions also perform robustly, achieving an F1-score of 0.889, which highlights the model's capacity to accurately detect beta-structured regions despite their lower prevalence. Coil regions, often more flexible and structurally diverse, yielded a slightly lower recall (0.818) but the highest precision (0.931), indicating that while the model tends to be conservative in predicting coil residues. The high F1-scores across all classes suggest that the model achieves a well-balanced trade-off between sensitivity and specificity. To further validate the effectiveness of the proposed UNet model, the Q3 accuracy metric was calculated. Using DSSP as the ground truth, and based on the estimated overlap of predicted and actual secondary structure assignments, the model achieved a Q3 accuracy of approximately 96%.

This high Q3 score indicates that the model accurately predicts the majority of residues, demonstrating reliable classification performance across structured (H, E) and unstructured (C) regions. The close agreement between the predicted secondary structure distributions and those reported by DSSP and STRIDE further supports the robustness and generalization ability of the UNet architecture.

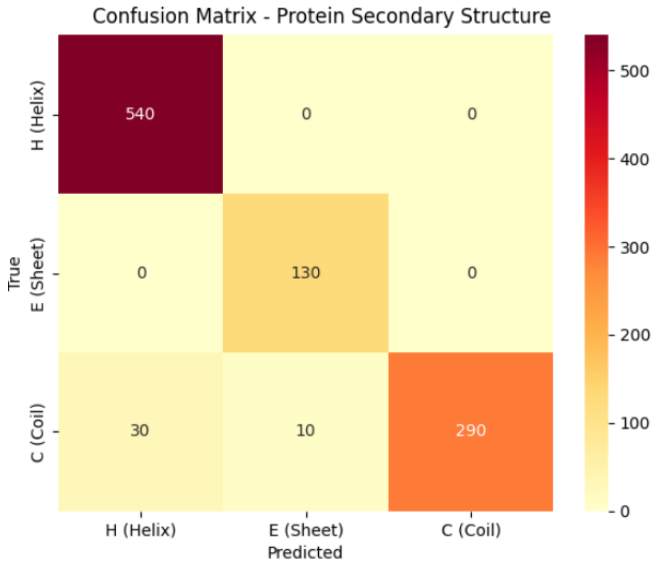

Figure 6. Confusion matrix.

The confusion matrix shows that the model perfectly classified Helix and Sheet structures, indicating strong learning of structured regions. Minor misclassifications occurred in Coil regions, which are typically harder to predict due to their unstructured nature. Overall, the model achieved a high Q3 accuracy of 96%, demonstrating excellent performance in secondary structure prediction.

To evaluate the secondary structure prediction capability of the proposed UNet-based model, the predicted content was compared with secondary structure elements derived from AlphaFold and RoseTTAFold predictions using DSSP and STRIDE assignments. As shown in **Table 2**, both DSSP and STRIDE reported approximately 54% alpha-helical content, with beta-sheet content at 13% and 11%, respectively. In comparison, the UNet model predicted 57% alpha helices and 14% beta sheets, closely aligning with the distributions obtained from structure-based tools. Additionally, the UNet model explicitly predicted 29% coil

regions, which were not directly quantified in the DSSP or STRIDE outputs shown. These results suggest that the proposed method effectively captures structural trends consistent with state-of-the-art predictors, with the added benefit of providing complete secondary structure classification from primary sequence alone.

Thus, beyond structural accuracy, the predicted secondary structure provides functional insights into QscR, a quorum-sensing transcriptional regulator in *Pseudomonas aeruginosa*. QscR belongs to the LuxR family of proteins, which typically exhibit a modular structure comprising an N-terminal ligand-binding domain and a C-terminal DNA-binding domain, often rich in α-helices. The high α-helical content (57%) predicted by the proposed UNet model aligns with the known helical architecture of these functional domains, particularly the helix-turn-helix motif critical for DNA recognition and binding. Additionally, the presence of well-defined β-sheets and coil regions suggests potential structural flexibility necessary for ligand interaction and allosteric regulation. By accurately mapping these structural elements, this findings contribute to a more detailed understanding of how QscR mediates quorum-sensing responses, potentially guiding future studies on inhibitor design or synthetic regulation of bacterial communication pathways.

7. Conclusion

In this study, a UNet based framework was developed and applied to predict the secondary structure of proteins, with a particular focus on the quorum-sensing repressor QscR from *Pseudomonas aeruginosa*. The model leveraged both sequence and evolutionary information through one-hot encoding and PSSMs to accurately classify residues into helix, sheet, and coil regions. The architecture's encoder-decoder structure with skip connections enabled it to capture both local and global dependencies effectively, resulting in strong predictive performance.The proposed UNet model achieved high precision, recall, and F1-scores across all secondary structure classes, and a Q3 accuracy of 96%, indicating robust generalization and reliability. Comparative evaluation with structure-based methods such as DSSP and STRIDE, as well as benchmark predictors like AlphaFold and RoseTTAFold, revealed that the predicted SS distribution closely matched experimentally derived and computationally inferred data. This validation confirms the model's effectiveness in real-world applications.The identification of prominent α-helices aligns with known DNA-binding domains, while β-sheet and coil predictions suggest additional flexibility and regulatory capability. These insights not only enhance molecular-level understanding of QscR but also highlight the potential of sequence-based deep learning approaches in functional annotation and drug discovery.

Future work will explore extending this approach to multi-sequence or multi-task learning models, integrating additional structural and functional annotations, and validating on broader protein families to further enhance predictive accuracy and biological relevance.

Funding statement

The authors received no specific funding for this study.

Conflicts of interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Author contributions statement

Conceptualization, Saravanan.K, Sivakumar.S; Methodology, Sangeetha.B, Marimuthu T; Investigation, Sivakumar.S ; Resources, Palanisamy P.N, Sangeetha P; Writing— Saravanan.K;

Data availability statement

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

1. Hondoh, T., Kato, A., Yokoyama, S. and Kuroda, Y. Computer-aided NMR assay for detecting natively folded structural domains. Protein science, 2006; 15(4), pp.871-883.
2. Shi, Q., Chen, W., Huang, S., Jin, F., Dong, Y., Wang, Y. and Xue, Z. DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network. Bioinformatics, 2019;35(24), pp.5128-5136.
3. Zheng, W., Zhou, X., Wuyun, Q., Pearce, R., Li, Y. and Zhang, Y. FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics, 2020;36(12), pp.3749-3757.
4. Guo, Z., Hou, J. and Cheng, J. DNSS2: Improved ab initio protein secondary structure prediction using advanced deep learning architectures. Proteins: Structure, Function, and Bioinformatics, 2021;89(2), pp.207-217.
5. Wu, T., Guo, Z., Hou, J. and Cheng, J. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. BMC bioinformatics, 2021; 22, pp.1-17.
6. Zhou, Y., Tan, K., Shen, X., He, Z. and Zheng, H, March. A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration. In 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE), 2024;(pp. 749-753). IEEE.
7. Cheng, J., Liu, Y. and Ma, Y. Protein secondary structure prediction based on integration of CNN and LSTM model. Journal of Visual Communication and Image Representation, 2020;71, p.102844.
8. Geethu, S. and Vimina, E.R. Protein secondary structure prediction using cascaded feature learning model. Applied Soft Computing, 2023; 140, p.110242.
9. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A. and Penedones, H. Improved protein structure prediction using potentials from deep learning. Nature, 2020; 577(7792), pp.706-710.
10. Roy, R.S., Quadir, F., Soltanikazemi, E. and Cheng, J. A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. Bioinformatics, 2022; 38(7), pp.1904-1910.
11. Jin, X., Guo, L., Jiang, Q., Wu, N. and Yao, S., 2022. Prediction of protein secondary structure based on an improved channel attention and multiscale convolution module. Frontiers in Bioengineering and Biotechnology, 10, p.901018.
12. Dong, B., Liu, Z., Xu, D., Hou, C., Dong, G., Zhang, T. and Wang, G., 2024. SERT-StructNet: Protein secondary structure prediction method based on multi-factor hybrid deep model. Computational and Structural Biotechnology Journal, 23, pp.1364-1375.
13. Dai, W., 2025. A Survey of Deep Learning Methods in Protein Bioinformatics and its Impact on Protein Design. arXiv preprint arXiv:2501.01477.
14. Ismi, D.P. and Pulungan, R. Deep learning for protein secondary structure prediction: Pre and post-AlphaFold. Computational and structural biotechnology journal, 2022;20, pp.6271-6286.
15. Stapor, K., Kotowski, K., Smolarczyk, T. and Roterman, I. Lightweight ProteinUnet2 network for protein secondary structure prediction: a step towards proper evaluation. BMC bioinformatics, 2022; 23(1), p.100.
16. Mahmud, S., Guo, Z., Quadir, F., Liu, J. and Cheng, J. Multi-head attention-based u-nets for predicting protein domain boundaries using 1d sequence features and 2d distance maps. BMC bioinformatics, 2022; 23(1), p.283.
17. Kabsch, W. and Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules, 22(12), pp.2577-2637.
18. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic acids research. 2004 Jul 1;32(suppl_2):W500-2.