International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(4):1103-1116 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i04.06682

Performance Analysis of Traditional Distance Metrics in Protein Structural Class Prediction

Shreya Saha[1], Papri Ghosh[1]*, Debmitra Ghosh[2], Subhram Das[1], Md Ashifuddin Mondal[1], Dharmpal Singh[2]

[1]Computer Science and Engineering, Narula Institute of Technology, Kolkata, India, [2]Computer Science and Engineering, JIS University, Kolkata, India. *Corresponding Author's Email: paprighosh06@gmail.com

Abstract

Predicting a protein's secondary structure directly from its amino acid sequence is a key challenge in bioinformatics. Successfully doing so has significant implications for understanding how proteins function and for designing new drugs. This study presents a comparative evaluation of seven distance and similarity measures—Euclidean, Manhattan, Minkowski, Cosine, Chebyshev, Mahalanobis, and Jaccard - for classifying proteins into four major secondary structural classes: α, β, α + β, and α/β. Using a curated dataset of 120 protein sequences represented by the frequency of 20 amino acids, each metric was employed in a minimum-distance-based classification framework. Group-wise frequency statistics, including mean, maximum, and minimum values, were analyzed to understand amino acid distribution across structural classes. A classification algorithm was then designed to compute distances between an unknown protein and each class group, identifying the closest match. Accuracy was measured by comparing predicted labels against true structural categories. The results show that the Mahalanobis distance achieved the highest mean classification accuracy (64.17%), closely followed by Cosine distance (61.67%), due to their ability to capture feature dependencies and directional similarity, respectively. Jaccard similarity performed poorly, indicating its inadequacy for continuous numerical data. The method yielded a maximum prediction accuracy of 79% for some cases. This comprehensive performance evaluation underscores the importance of selecting appropriate distance metrics for structural classification tasks and sets the foundation for future integration with ensemble or deep learning models.

Keywords: Amino Acid Frequency, Bioinformatics, Distance Metrics, Prediction Models, Protein Secondary Structure (PSS), Secondary Structural Classes (SSC).

Introduction

The function of proteins, the essential bricks of biological systems, is closely tied to their intricate 3D structures (1). The way these proteins are organized into secondary structure classes—specifically α-helices, β-sheets, and combinations of the two—is a key factor in determining their biological roles (2). For years, a major focus in computational biology and bioinformatics has been predicting these secondary structure classes based on a protein's amino acid sequence. This ability provides vital information for annotating protein function, discovering new drugs, and gaining insights into disease mechanisms (3). Traditional approaches for predicting protein secondary structures (PSS) have relied on statistical methods, machine learning algorithms, and sequence alignment techniques. Recently, distance and similarity-based models have gained attention due to their simplicity, interpretability, and effectiveness, particularly when the structural characteristics are embedded in amino acid composition profiles. Selecting an appropriate distance measure becomes critical because it directly impacts the ability to capture subtle variations and patterns in amino acid frequency distributions across protein classes. In the literature, several efforts have been made to predict protein structure classes using amino acid compositions. For instance, Chou introduced early statistical models in a 19-dimensional composition space, demonstrating the predictive value of amino acid frequencies (4). More recent studies have incorporated various distance functions to enhance prediction accuracy, yet a systematic comparison across multiple distance metrics remains limited. In this research, systematically performance have been evaluated of seven widely used distance and similarity measures- Euclidean,

(Received 04th July 2025; Accepted 23rd September 2025; Published 30th October 2025)

Manhattan, Minkowski, Cosine Distance, Chebyshev, Mahalanobis, and Jaccard similarity—for classifying proteins into their secondary structure classes based solely on amino acid sequence information. A dataset containing 120 proteins classified into α, β, α + β, and α/β categories is utilized, and the classification is performed using a minimum distance-based strategy. An overview of the methodology is presented in Figure 1, illustrating the key steps: data preprocessing, distance computation, protein class prediction, and performance evaluation.

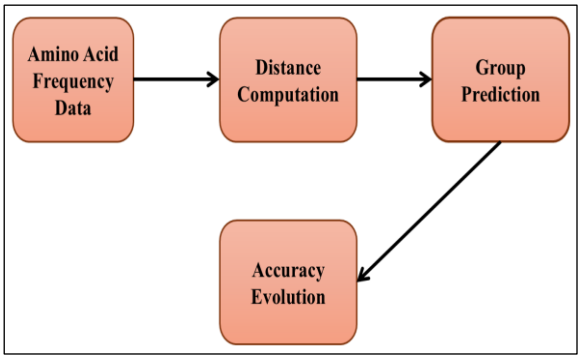

Figure 1: Overview of the Protein Structure Classification Process Based on Distance Metrics

This work offers several methodological contributions that distinguish it from existing approaches in protein structural class prediction. First, the dataset of 120 proteins was carefully curated to ensure representation across all four major secondary structural classes (α, β, α+β, α/β). This balanced dataset allows for fairer comparison of classification methods compared to commonly used imbalanced benchmarks.

Second, each protein sequence was transformed into a 20-dimensional amino acid frequency vector. This serves as a form of dimensionality reduction, compressing variable-length sequences into fixed-length representations while preserving essential biochemical information. Such a representation not only simplifies computation but also provides biological interpretability, as class-specific amino acid usage can be directly analyzed through mean, maximum, and minimum frequency statistics.

Third, unlike previous studies that applied individual distance measures in isolation, this study presents a systematic and comparative evaluation of seven widely used distance and similarity metrics under a uniform minimum-distance-based classification framework. The results reveal that Mahalanobis and Cosine distances best capture biological and statistical dependencies in amino acid composition, highlighting their interpretability in terms of correlation and directional similarity.

Together, these aspects—dataset curation, interpretable dimensional reduction, and systematic comparative evaluation—form the unique contributions of this work to protein structural class prediction.

The remainder of the paper contains the discussion on related work and background studies, the materials and methods, results and discussion, and the conclusion of the study with insights and future work.

The analysis of genome sequences is fundamental to understanding evolutionary relationships, genetic variation, and functional genomics across species. Traditional alignment-based methods, though useful, become inefficient when dealing with large and complex datasets. Recent advancements, such as the use of alignment-free techniques based on numerical descriptors and distance metrics, allow for faster and more scalable comparisons (5, 6). These methods convert nucleotide sequences into multidimensional numeric vectors and apply metrics like Bray-Curtis or positional difference-based descriptors to construct accurate phylogenetic trees. Such strategies not only improve the efficiency of genome sequence analysis but also enhance the accuracy of identifying evolutionary links, even among datasets of unequal lengths. This opens up new avenues in comparative genomics, taxonomy, and evolutionary biology.

In parallel, protein sequence and structure analysis perform a vital role in deciphering the functional and structural dynamics of biological molecules. The alignments of the protein-protein

interaction (PPI) network and secondary structure comparisons provide information on the structural similarities that govern protein function (7, 8). Understanding these structural elements allows researchers to predict protein functions, identify conserved motifs, and explore cross-species functional orthologs. Novel approaches like the TOPS-based numerical descriptors and moment-of-inertia analysis demonstrate superior performance in generating phylogenetic trees for structural classification. As proteins are the primary effectors of cellular function, these techniques are pivotal in drug discovery, functional annotation, and systems biology, ultimately bridging the gap between sequence data and biological insights.

Grasping a protein's structure is essential for understanding how it functions, which in turn is crucial for progress in fields like molecular biology, pharmacology, and medicine. Unfortunately, current experimental methods for determining these structures, such as X-ray crystallography and NMR spectroscopy, are expensive, require significant time and effort, and are often very slow (4). As a result, computational approaches have emerged as efficient and scalable alternatives that complement traditional techniques by accelerating structural prediction and functional annotation.

Recent breakthroughs in machine learning have pointedly enhanced the accuracy of protein structure prediction. It is emphasized by Seok et al. that accurate prediction of secondary structures is important for measuring the three-dimensional (3D) conformation of a protein, which is critical for understanding molecular functions, protein–protein interactions, and disease mechanisms (9). Tools such as AlphaFold and RoseTTAFold demonstrate remarkable performance in modeling complex protein structures, thereby enhancing disease modeling and drug discovery pipelines.

The transformative impact of deep learning-based models such as AlphaFold2, which enable large-scale and high-throughput structural predictions, is further highlighted (10). These tools have played a pivotal role in research areas such as vaccine design, mutation effect analysis, and protein engineering. Moreover, emerging models such as ESMFold are addressing the challenges of predicting the structural consequences of sequence variations, contributing to a deeper understanding of protein dynamics.

The value of accurate structure prediction is reinforced, with the determination of protein structure asserted by AlQuraishi et al. as essential for uncovering biological function, facilitating drug discovery, and supporting protein engineering efforts (11). Advances in deep learning and machine learning have significantly enhanced structural modeling, bridging the gap between sequence data and functional understanding.

The importance of secondary structure prediction is also echoed in the work of Pakhrin et al., where its role in understanding protein function, cellular processes, and interaction networks is noted (12). With the growing gap between rapidly accumulating protein sequences and experimentally resolved structures, computational methods have become increasingly vital for research in drug design and the study of protein misfolding diseases.

The view that secondary structure prediction serves as a foundational step in protein function annotation is emphasized in the study that depicts that accurate prediction provides crucial insights into 3D conformation and disease mechanisms (13). By introducing improved models like MLPRNN, the researchers aim to make protein structure prediction more efficient and scalable, particularly for large and complex datasets.

The limitations of alignment-based methods for structure prediction, particularly in handling large datasets or orphan proteins, are addressed in a study where the work is on the RGN2 model, which predicts structures from single sequences using alignment-free techniques, demonstrates the potential of such approaches in functional annotation, novel protein design, and systems biology (14).

An alignment-free, embedding-based method using protein language models to efficiently predict conservation patterns was proposed by where unlike traditional alignment-based approaches, which are computationally intensive and sensitive to sequence order, their method enables accurate conservation 4 analysis even in multi-domain or fast-evolving proteins (15). This has important applications in identifying functional domains, regulatory elements, and in accelerating drug discovery.

The narrowing of the conformational search space and the improvement of functional inference for low-homology sequences are achieved through

structural class prediction is also highlighted (16). The proposed method enhances prediction accuracy, contributing to bioinformatics research in drug discovery, functional annotation, and systems biology.

Studies Based on Identical Dataset

The dataset selected for this research has been widely used by many other researcher, and it is obtained from and also been utilized by several other researcher (4). The structure of the insulin molecule was investigated using a novel cybernetic and mathematical model is approached (17). By analyzing the atomic composition of amino acids and calculating standard deviations, the study revealed a digital bio-code underlying insulin's sequence. The findings suggest a balance of positive and negative deviations, indicating a possible algorithmic structure in protein biochemistry. This approach opens new pathways for digital modeling in bioinformatics and genetics.

An alignment-free method for identifying Soluble N-ethylmaleimide - Sensitive factor Attachment Protein Receptor (SNARE) proteins was introduced by utilizing multi scan convolutional neural networks (CNNs) based on PSSM profiles (18). Protein sequences were encoded into 20×20 matrices to capture evolutionary data, and SMOTE was applied to counter class imbalance. The CNN, equipped with varying filter sizes, extracted discriminative features effectively, reaching 95.5% accuracy and an AUC of 0.963. t-SNE and UMAP were used for visualization, confirming a clear separation between SNARE and non-SNARE classes.

A two-level multi-label classification system known as iAMP-2L has been introduced to recognize antimicrobial peptides (AMPs) and determine their functional roles. The classifier initially detects whether a peptide is an AMP and subsequently classifies its functional categories—even when they belong to multiple types. Using a combination of pseudo-amino acid composition (PseAAC) and a fuzzy K-nearest neighbour (FKNN) algorithm, this model achieves high accuracy in classification. The tool is easily accessible through a user-friendly web server. Significant promise for advancing antimicrobial drug discovery through accurate peptide function prediction is shown by this approach (19).

An automatic feature learning framework for activity recognition was proposed which leveraged Principal Component Analysis (PCA) and deep autoencoders on raw sensor data (20). To preserve structural patterns, ECDF-based normalization was introduced. Evaluations across four public datasets demonstrated that learned features, especially from deep models, consistently outperformed handcrafted ones, even under sparse data conditions.

The development of a neural network–based method to determine protein subcellular location using only amino acid composition is attributed which didn't depend on similar sequences or motifs, achieved an accuracy of 81% for prokaryotic proteins and 66% for eukaryotic proteins (21). The method demonstrated robustness even when the first few amino acids were incorrect and maintained reliable performance on independent data. This prediction tool was made obtainable, facilitating genome analysis and protein location prediction.

The influence of the hydrophobic effect on protein interactions was examined through an extensive statistical analysis was performed that involved 362 protein-protein interfaces and 57 oligomeric interfaces (22). They measured hydrophobicity based on factors like amino acid composition, interactions between amino acid residues, and the amount of buried nonpolar surface area. Although the hydrophobic effect was found to be significant in protein binding, it was less dominant than in folding. Charged and polar residues appeared to contribute more to interface stability. The study underscored distinctions between monomer folding and protein-protein interactions, especially for interface-specific model design.

The introduction of DPP-PseAAC, a computational model for predicting DNA-binding proteins based solely on amino acid sequence, was reported which involved extracting features using Chou's general PseAAC, then using Random Forest to rank these features (23). Finally, they trained the model with an SVM (linear kernel), and a process called Recursive Feature Elimination (RFE). The method showed superior performance compared to existing predictors and was made publicly accessible via a web server.

A predictive framework for drug–target interaction was developed by encoding drugs based on their functional group compositions and proteins using biologically meaningful features. Using mRMR for feature selection followed by

Nearest Neighbor classification, the model achieved an accuracy exceeding 78% across four distinct protein families. Demonstrating both efficiency and strong predictive capability, this approach offers practical value in drug discovery efforts (24). A detailed summary of the survey's findings, including the study focus, methodology, performance/results, and application area for each work, is presented in Table 1.

Table 1: Summary of Studies Involving Machine Learning and Statistical Methods in Protein and Bioinformatics Research

Study	Focus	Methodology	Performance / Results	Application Area
Kuri *et al*. (17)	Structure of insulin	Cybernetic and mathematical analysis using atomic composition and standard deviation	Revealed digital bio-code, balanced deviations	Bioinformatics, protein
Kha *et al*. (2022) (18)	SNARE protein detection	CNN model with PSSM profiles, SMOTE, and dimensionality reduction (t-SNE, UMAP)	95.6% accuracy, AUC 0.963	Protein classification
Xiao *et al*. (2013) (19)	Antimicrobial peptide classification	Two-level multi-label classification using SVM and kNN with web server integration	Multi-label AMP function prediction	Drug discovery, peptide classification
Plötz *et al*. (2011) (20)	Activity recognition	Feature learning with PCA, deep autoencoders, and ECDF normalization	Outperformed handcrafted features	Sensor data analysis, pattern recognition
Reinhardt *et al*. (1998) (21)	Protein subcellular localization	Neural network model using amino acid composition without homology	81% (prokaryotes), 66% (eukaryotes)	Genome analysis, protein targeting
Tsai *et al*. (1997) (22)	Protein-protein interface study	Statistical analysis of hydrophobicity and residue interaction	Charged/polar residues stabilize interface	Protein interface
Rahman *et al*. (2018) (23)	DNA-binding protein prediction	Sequence-based Machine Learning (ML) using Chou's Random Forest and SVM with recursive feature elimination	Outperformed existing models	Genomic analysis, protein-DNA interaction
He *et al*. (2010) (24)	Drug-target interaction prediction	ML framework using QSOAR feature selection and Nearest Neighbour classifiers	>78% accuracy	Drug discovery, pharmacological prediction

Methodology

The study classifies 120 proteins into four main structural categories: Alpha (α), Alpha+Beta (α+β), Beta (β), and Alpha/Beta (α/β). This grouping is based on the dominant type of secondary structure in each protein. The dataset, used for this classification, includes the frequencies of the 20 different amino acids for each of the 120 proteins (4, 25).

The α class contains proteins primarily made of α helices, while the β class is dominated by β sheets. Proteins in the α+β class have distinct, separate sections of α helices and β sheets. In contrast, the α/β class is characterized by an intricate mix of α helices and β sheets woven together throughout the protein's structure. This system of classification is useful for understanding how proteins are organized and for identifying functional similarities between them. The dataset's amino acid frequency information can also be used to predict the names of the proteins.

Group Frequency Measure

To further analyze the dataset, we measured the group frequencies of all the 20 amino acid groups based on their occurrence across the four structural classes. Specifically, to calculate the

amino acid of each group (α, α/β, α+β, and β), we consider the Average Values of amino acids across different protein structure groups (summarized in Table 2), Maximum Value of amino acids across different protein structure groups (presented in Table 3), and Minimum Value (shown in Table 4) of the entire amino acid frequencies. The algorithm for this is as provided in Algorithm 1.

Algorithm 1: Computation of Group-Level Amino Acid Frequencies

Input: A matrix F of size 20 × n, where F_{ij} represents the frequency of the i^{th} amino acid in the j-th species of a group.

Output: A vector G of size 20, where G_i is the mean frequency of the i^{th} amino acid for the group.

Procedure

1. **for** i = 1 to 20 do ▷ Iterate over each amino acid
   a. $G_i = \frac{1}{n}\sum_{j=1}^{n} F_{ij}$ ▷ Compute Mean Frequencies
2. **end for**
3. **Return** G

In Step.1 of Algorithm.1, the equation for mean frequency calculation is portrayed. We may use the following two equations, Eq.1 and Eq.2 to obtain the frequencies via max or min values respectively.

$$G_i \leftarrow F_{ij} \quad \text{[Eq.1]}$$

$$G_i \leftarrow F_{ij} \quad \text{[Eq.2]}$$

Table 2: Mean Values of Amino Acids across Different Protein Structure Groups

Group	α	α/β	α+β	β
A	11.05983	9.246	6	6
C	0.969533	1.106667	2.786667	2.786667
D	5.553	5.066	4.968333	4.968333
E	7.460583	6.169	4.976667	4.976667
F	3.976667	3.296667	4.982667	4.982667
G	6.206833	8.083167	7.509033	7.50903
H	1.0125	2.127	1.414	1.414
I	4.025	6.111667	5.054667	5.054667
K	8.59	6.316	6.122	6.122
L	11.27267	7.655833	7.018667	7.018667
M	2.50125	3.156167	1.815167	1.815167
N	2.392	4.311667	5.131667	5.131667
P	4.993	5.651667	5.84	5.84
Q	4.321	4.058	4.296967	4.296967
R	4.584	4.308	3.855	3.855
S	5.136	5.546	8.078667	8.078667
T	5.432833	5.286667	7.672	7.672
V	5.461583	5.8775	6.704333	6.704333
W	4.021833	1.55	1.583	1.583
Y	2.944	6.262667	4.344	4.344

The average values provide an overall estimate of the typical amino acid composition across proteins within a given structural class, offering insight into common trends and characteristic patterns. The maximum values highlight the amino acids that are most frequently occurring within each group, identifying key residues that may play critical roles in maintaining the structural integrity of proteins in that class. Conversely, the minimum values reveal the least occurring amino acids, suggesting

potential differences in amino acid utilization based on structural constraints. This comprehensive statistical assessment enables a deeper understanding of how amino acid composition varies among different protein structural organizations and can further assist in predictive modeling of protein classes based on sequence information.

Table 3: Max Values of Amino Acids across Different Protein Structure Groups

Group	α	α/β	α+β	β
A	22.05	17.72	18.69	18.69
C	9.23	2.72	20	20
D	11.32	10.88	11.24	11.24
E	16.13	13.77	11.65	11.65
F	10.29	8.05	19.35	19.35
G	13.61	12.24	16.16	16.16
H	8.5	4.76	6.45	6.45
I	9.76	10.87	12.12	12.12
K	16.13	10.61	16.67	16.67
L	19.35	12.26	13.16	13.16
M	7.14	4.17	5.41	5.41
N	7.69	7.94	12.9	12.9
P	7.14	9.52	13.51	13.51
Q	10.74	7.1	12.15	12.15
R	15.79	7.29	13.33	13.33
S	9.88	13.82	14.29	14.29
T	8.57	9.04	16.53	16.53
V	12.33	17.46	11.21	11.21
W	4.52	2.97	3.94	3.94
Y	9.76	9.52	11.36	11.36

Table 4: Min Values of Amino Acids across Different Protein Structure Groups

Group	α	α/β	α+β	β
A	0	1.59	0	0
C	0	0	0	0
D	0	3.17	0	0
E	2.04	1.45	0.93	0.83
F	1.09	0	0	0
G	1.75	4.84	2.02	2.02
H	0	0	0	0
I	0	0	0.88	0.88
K	2.01	2.72	0	0
L	3.08	5	3.03	3.03
M	0	0	0	0
N	0	2.04	0	0
P	0	1.75	1.18	1.18
Q	0.56	1.45	0	0
R	0.65	0.73	0	0
S	0	1.52	0	0

Group	α	α/β	α+β	β
T	0	1.52	0	0
V	0	4.52	0	0
W	0	0	0	0
Y	0	1.11	0	0

Distance Measure

To analyze the performance on predicting Protein Secondary Structure Classes from Amino Acid Sequences, seven traditional distance and similarity measures have been considered.

The Euclidean Distance (Eq.3) measures the direct longitudinal distance between two points and is commonly applied in continuous space analysis (26).

$$d(\alpha, \beta) = \sqrt{(\alpha - \beta)^2} \quad \text{[Eq.3]}$$

The Manhattan Distance (Eq.4) measures the sum of the absolute gaps of their coordinates, often applied when movement is restricted to grid-like paths (26).

$$d(p, q) = \sum_{i=1}^{n} |p_i - q_i| \quad \text{[Eq.4]}$$

The Minkowski Distance (Eq.5) simplifies both Euclidean and Manhattan distances by introducing a parameter p, allowing flexibility in the measurement scale (26).

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad \text{[Eq.5]}$$

The Cosine Distance (Eq.6) assesses the angular difference between two vectors, focusing on their orientation rather than magnitude, and is particularly useful in text and high-dimensional data analysis (26).

$$d_{cos}(x, y) = 1 - \frac{\sum_{i=0}^{n-1} x_i y_i}{\sqrt{\sum_{i=0}^{n-1} x_i^2} \times \sqrt{\sum_{i=0}^{n-1} y_i^2}} \quad \text{[Eq.6]}$$

The Mahalanobis Distance (Eq.7) measures the distance between points while considering correlations among variables, making it suitable for multivariate data (27).

$$d(p_m, p_n) = \sqrt{(p_m - p_n)^\top A (p_m - p_n)} \quad \text{[Eq.7]}$$

The Chebyshev Distance (Eq.8) captures the maximum absolute difference among corresponding vector components, highlighting the dominant difference (26).

$$dist(A, B) = (|x_A - x_B|, |y_A - y_B|) \quad \text{[Eq.8]}$$

Finally, the Jaccard Similarity Coefficient (Eq.9) measures the similarity between finite sample sets, widely used for binary and set-based data comparisons (26).

$$jac(p, q) = \frac{|p \cap q|}{|p| + |q| - |p \cap q|} \quad \text{[Eq.9]}$$

Protein Species Group Selection

After calculating the distance measure between a selected species and each of the 19 reference species, the analysis is performed based on four distinct protein sequence categories: α, β, α + β, and α/β. For each category, the distances between the selected species and the species within that category are computed separately. The minimum distance for each group is then determined.

Algorithm 2 Protein Structure Group Classification Using Minimum Distance Criterion

Procedure PredictProteinClass(P, C_α, C_β, $C_{\alpha+\beta}$, $C_{\alpha/\beta}$)

1. classes ← {C_α, C_β, $C_{\alpha+\beta}$, $C_{\alpha/\beta}$}
2. minimum distances ← [ ]
   ▷ List to hold the minimum distance for each class
3. For each C in classes do
   i. temp distances ← [ ]
   ii. For each sample in C do
      1. distance ← ComputeDistance(P, sample)
      2. Append distance to temp distances
   iii. End For
   iv. min distance ← min(temp distances)
   v. Append min distance to minimum distances
4. End For
5. predicted index ← arg min(minimum distances)
6. Return corresponding class from classes[predicted index]

Ultimately, the group with the smallest minimum distance among the four categories is considered the most similar to the selected species. Therefore, the species is classified into that group, as it shares the most similar sequence characteristics according to the distance metric. The procedure for determining the correct group is outlined in Algorithm 2.

The algorithm has O(km) time complexity, where k is protein classes and m is samples in each class. This is because for each class, the algorithm computes the distance for each sample, and the distance computation is done for all classes.

Accuracy Measure

To evaluate the performance of the proposed classification approach, an accuracy assessment algorithm (Algorithm 3) is used. This algorithm receives two input lists: the true group labels and the predicted group labels for the 19 protein species. It evaluates each actual–predicted label pair and records the count of accurate predictions. The accuracy is then measured as the proportion of correctly classified species to the total number of species, with the result multiplied by 100 to represent it as a percentage. This straightforward yet effective metric provides a clear indication of the model's effectiveness in correctly categorizing protein sequences into their corresponding structural groups.

The algorithm has a time complexity of O(n), with *n* representing the total number of protein species (or the length of the input lists). This is because the algorithm iterates through each pair of actual and predicted labels exactly once.

Algorithm 3 Computation of Classification Accuracy for Protein Groups

Procedure ComputeClassificationAccuracy(TrueLabels, PredictedLabels)

1. correctCount ← 0
2. numSamples ← length(T rueLabels)
3. **For** i = 1 to numSamples do
   a. **If** T rueLabels[i] = P redictedLabels[i] then
      i. correctCount ← correctCount + 1
   b. **End If**
4. **End For**
5. accuracyPercentage ← $\frac{correctCount}{numSamples} \times 100$
6. **Return** accuracyPercentage

Results and Discussion

Table 5 shows the performance assessment results of seven distance metrics used for predicting protein secondary structure classes from amino acid sequences. Each distance measure—Euclidean, Manhattan, Minkowski, Cosine Distance, Chebyshev, Mahalanobis, and Jaccard—has been assessed based on its minimum, mean, and maximum accuracy values. Among these, the Mahalanobis distance achieved the highest mean

accuracy (64.17%), highlighting its ability to capture inter-variable relationships critical for protein classification tasks. The Cosine distance also performed competitively with a mean accuracy of 61.67%, whereas the Jaccard similarity consistently recorded the lowest performance (25%), indicating its unsuitability for numerical frequency-based data. Additionally, a visual comparison of these performance metrics is shown in Figure 2, where it is evident that Mahalanobis and Cosine distances outperform others. The combined analysis from Table 5 and Figure 2 underscores the importance of selecting appropriate distance metrics to augment the performance of accuracy of protein secondary structure prediction models.

Table 5: Accuracy Measure of all Distance Values

Distance Measure	Minimum	Mean or Average	Maximum
Euclidean	25	63	25
Manhattan	25	58.34	25
Minkowski	25	60.84	25
Cosine Distance	27.5	61.67	42.5
Chebyshev	25.83	55	29.17
Mahala Nobis	25	**64.16**	25
Jaccard	25	25	25

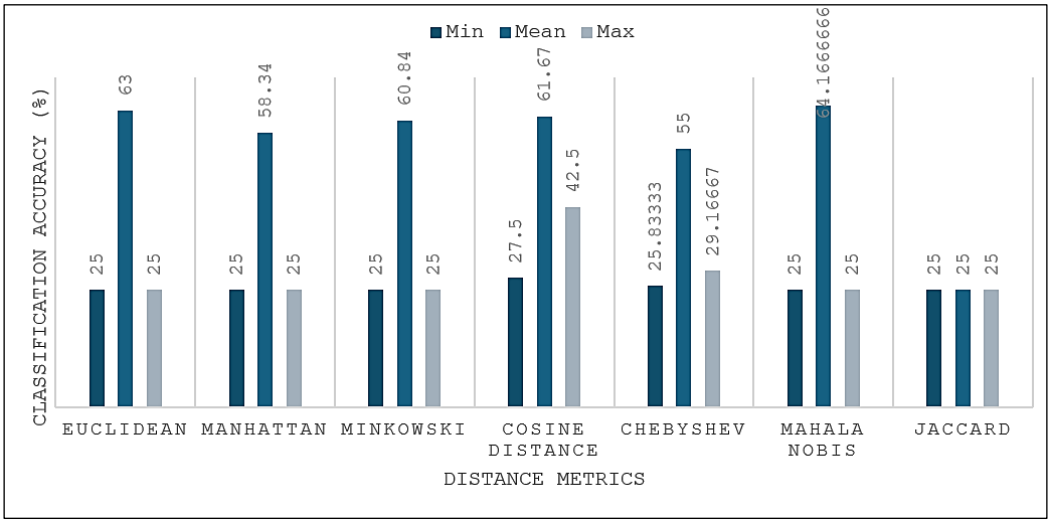

Figure 2: Graphical Representation of Accuracy for Seven Distance Measure

Validation

Process 1: Two datasets consisting of 86 and 106 protein sequences are collected from Protein Data Bank (PDB) (28). Table 6 presents 86 proteins representing three distinct families: α, β, and α + β as described in the Sierk-Pearson database (29). Table 7 classifies 106 proteins into four unique taxonomies: α, β, α+β, α/β (4).

Table 6: Classification of 86 Proteins by Sierk-Pearson

α	α + β	β
1ad6, 1a06, 1bbh, 1cns, 1d2z, 1dat, 1e12, 1eqz, 1gwx, 1hgu, 1hlm, 1jnk, 1mmo, 1nub, 1quu, 1rep, 1sw6, 1trr, 2hpd, 2mta	1a8d, 1a8h, 1aoz, 1b8m, 1bf2, 1bjq, 1bqy, 1btk, 1c1z, 1cl7, 1d3s, 1dan, 1dsy, 1dxm, 1et6, 1ext, 1nfi, 1nuk, 1otc, 1qdm, 1qe6, 1qfk, 1que, 1rmg, 1tmo, 2tbv	1a1m, 1a2v, 1akn, 1aqz, 1asy, 1ati, 1auq, 1ax4, 1b0p, 1b2r, 1bcg, 1bcm, 1bf5, 1bkc, 1bp7, 1c4k, 1cd2, 1cdg, 1d0n, 1d4o, 1d7o, 1doi, 1dy0, 1e2k, 1ecc, 1fbn, 1gso, 1mpy, 1obr, 1pty, 1qb7, 1qmv, 1urn, 1zfj, 2acy, 2drp, 2nmt, 2reb, 4mdh, 5uoj

Table 7: Classification of 106 Proteins

α	α + β	β	α / β
1avh, 1bab, 1brd, 1c5a, 1cpc, 1eco, 1fcs, 1fha, 1fia, 1hbg, 1hdd, 1hig, 1le4, 1lts, 1mbc, 1rpr, 1tro, 1utg, 256b, 2ccy, 2lh1, 2lhb, 2mhb, 2zta, 4mba, 4mbn	2aak, 1ctf, 1dnk, 1eaf, 1hsb, 1lts, 1oia, 1poc, 1ppn, 1rnd, 1snc, 1tfg, 1tgs, 2ach, 2act, 2bpa, 2sns, 3ssi, 3il8, 3rub, 3sgb, 3sic, 4blm, 4tms, 8cat, 9rnt, 9rsa	41acx, 2ayh, 1cd8, 1cdt, 1cid, 1dfn, 1hil, 1hle, 1mam, 3mon, 2phy, 1rei, 1ten, 1tlk, 2vaa, 2alp, 2avi, 2bpa, 3hhr, 2ila, 2lal, 2snv, 3cd4, 4gcr, 7api, 8fab	1aba, 1cis, 1cse, 1dhr, 2dri, 1etu, 1fx1, 1gpb, 1pax, 1pfk, 2pgd, 1q21, 1s01, 1sbp, 1sbt, 1tim, 1tre, 1ula, 1bks, 2had, 2liv, 3gbp, 2fox, 4cpa, 5p21, 8abp, 8atc

Table 8: Accuracy Measures of all Distance Values of 86 and 106 Dataset

Distance Measure	Accuracy on 86 Dataset	Accuracy on 106 Dataset
Euclidean	70.33	65.48
Manhattan	66.67	63.6
Minkowski	62	60.5
Cosine Distance	67.5	63
Chebyshev	56	55
Mahala Nobis	72	74.16
Jaccard	29	26

In both datasets presented in Table 8, a consistent technique has been employed to predict the group classification of a selected protein. Specifically, the method utilizes the mean value to represent the group frequency, as this approach has demonstrated favorable performance, as evidenced in Table 5. The rationale behind using the mean value lies in its ability to provide a stable and representative central tendency, which contributes to improved prediction accuracy. To assess the efficiency of the proposed method, we have assessed the performance of several distance-based similarity measures, including Euclidean, Manhattan, Minkowski, Cosine, Chebyshev, Mahalanobis, and Jaccard distances. These metrics were systematically applied to both datasets to determine their impact on classification performance. The resulting outcomes highlight the comparative strengths of each distance measure under the given experimental conditions.

Process 2: The dataset of 120 protein sequences, represented by the frequency of 20 amino acids and categorized into four structural classes—Alpha (α), Beta (β), Alpha+Beta (α+β), and Alpha/Beta (α/β)—was analyzed using seven distance metrics, where Mahalanobis distance stands out the highest mean accuracy (64.17%) and to further validate the findings, the same dataset was reanalyzed using both distance-to-reference and k-NN classifiers (k=3,5) (4). The use of k-NN validated the effectiveness of distance metrics within a standard classifier, confirming that the observed performance trends were not restricted to the distance-to-reference approach. A graphical representation and classification accuracy of these performance metrics are shown Figure 3 and Table 9.

Table 9: Accuracy (%) of Distance-Based and k-NN Classification for Seven Distance Metrics

Metric	Distance-to-Mean	Distance-to-Min	Distance-to-Max	k-NN (k=3)	k-NN (k=5)
Euclidean	62.50%	25.00%	25.00%	60.00%	59.17%
Manhattan	58.33%	25.00%	25.00%	56.67%	55.83%
Minkowski	62.50%	25.00%	25.00%	60.00%	59.17%
Cosine	61.67%	27.50%	42.50%	63.33%	60.83%
Chebyshev	55.00%	25.83%	29.17%	60.83%	60.00%
Mahalanobis	64.17%	25.00%	25.00%	54.17%	55.00%
Jaccard	25.00%	25.00%	25.00%	32.50%	32.50%

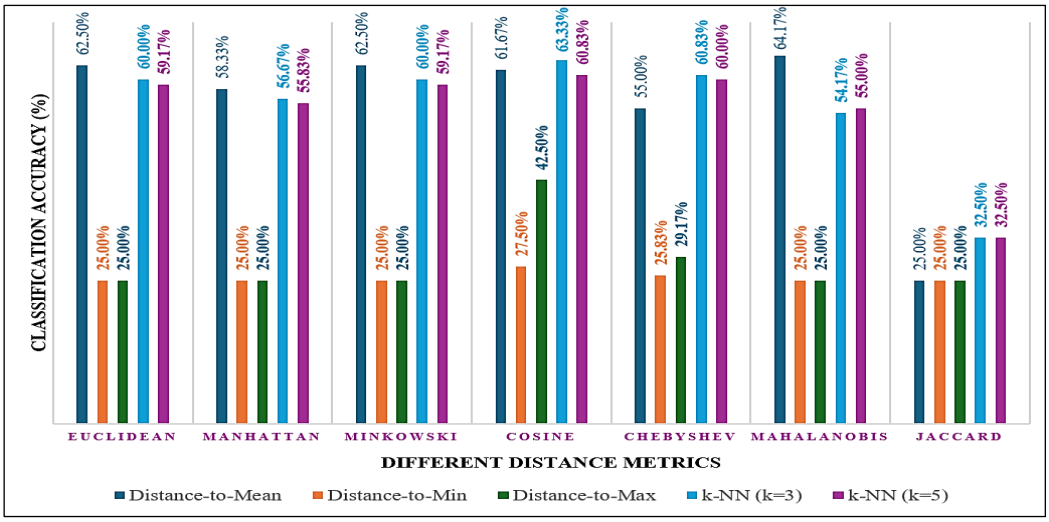

Figure 3: Graphical Representation of Distance-Based and k-NN Classification for Seven Distance Metrics

Table 9 represents the classification of accuracy (%) of seven distance metrics for protein secondary structure prediction using both distance-to-reference and k-NN classifiers (k=3 and k=5). Distance-to-Mean consistently generated the highest accuracies, with Mahalanobis distance performing best at 64.17%, Closely approached by Euclidean and Minkowski (62.5%) and Cosine (61.67%). Distance-to-Min and Distance-to-Max profiles resulted in much lower accuracies, mostly between 25% and 42%, indicating that extreme values are less representative of structural classes. In the k-NN classification, similar trends were observed: Cosine distance achieved slightly higher accuracy (63.33% for k=3), Mahalanobis showed lower performance (54.17–55%), and Jaccard remained unsuitable for this dataset (32.5%). Overall, the validation confirms that mean-based reference profiles are the most representative and effective for distance-based protein classification with Mahalanobis demonstrated the highest accuracy for protein secondary structure prediction.

The accuracy obtained with Mahalanobis distance (64.17%) is lower than recent deep learning-based approaches such as CNN-based SNARE detection, which reported 95.6% accuracy (18). However, deep learning models are computationally intensive and require large curated datasets, whereas the present approach offers a lightweight and interpretable solution suitable for scenarios with limited resources. The reliance on amino acid composition and sequence-based comparisons in our study aligns with earlier investigations where homology modeling and sequence similarity were effectively used to explore evolutionary relationships, such as in the analysis of the LFY gene across plant families (30). These findings support the utility of alignment-free and distance-based approaches in revealing biologically meaningful patterns, even when predictive accuracy is moderate.

Conclusion

In this study, we evaluated the effectiveness of seven traditional distance and similarity measures - Euclidean, Manhattan, Minkowski, Cosine, Chebyshev, Mahalanobis, and Jaccard for classifying protein secondary structure types based on amino acid sequence data. By analyzing a curated dataset of 120 proteins categorized into four major secondary structure classes, we performed a systematic comparison of these metrics using a minimum distance-based classification approach.

Our results demonstrate that the Mahalanobis distance metric achieves the highest average classification accuracy (64.17%), closely followed by the Cosine distance (61.67%), indicating their superior ability to capture important patterns and relationships among amino acid frequencies. In contrast, the Jaccard similarity consistently showed the lowest performance, highlighting its limitation in handling continuous frequency data in this domain.

The study also underlines the significance of choosing appropriate distance measures to enhance predictive performance in bioinformatics applications. Through the proposed methodology, a maximum observed prediction accuracy of 79% was achieved, suggesting the potential of

information-based and correlation-aware distance measures in improving protein structure classification tasks.

Future work could extend this approach by integrating ensemble techniques that combine multiple distance metrics or hybrid similarity measures, thereby enhancing predictive robustness and accuracy across diverse protein families. In addition, incorporating deep learning-based embeddings and evaluating performance on larger, more diverse datasets from repositories such as the Protein Data Bank (PDB) would further validate and generalize the findings. Beyond structural class prediction, the insights gained here may also be applied to related bioinformatics challenges, including drug–target interaction studies, protein functional annotation, and the classification of intrinsically disordered proteins, where accurate sequence–structure modeling plays a pivotal role.

Abbreviations

AMPs: Antimicrobial Peptides, DNA: Deoxyribonucleic Acid, ML: Machine Learning, PCA: Principal Component Analysis. PDB: Protein Data Bank, PPI: Protein-Protein Interaction, PSS: Protein Secondary Structure, SNARE: Soluble N-ethylmaleimide – Sensitive factor Attachment Protein Receptor.

Acknowledgment

We wish to express our sincere gratitude to Narula Institute of Technology, Kolkata, India and JIS University for their support and research environment that contributed to the successful completion of this study. We would also like to acknowledge the valuable contributions of all the co-authors - Shreya Saha, Debmitra Ghosh, Subhram Das, Md. Ashifuddin Mondal and Dharmpal Singh - whose collaborative efforts were instrumental in developing the methodologies and insights presented in the paper titled "Efficient DNA Sequence Classification through Machine-Learning Techniques."

Author Contributions

Shreya Saha: Conceptualization, Methodology, Data Curation and Validation, Papri Ghosh: Conceptualization, Methodology, Data Curation and Validation, Overall supervision, Subhram Das: Conceptualization, Methodology, Data Curation and Validation, Overall supervision, Debmitra Ghosh: Data Curation and Validation, Md. Ashifuddin Mondal: Overall supervision, Dharmpal Singh: Overall supervision.

Conflict of Interest

The authors of this work state that they have no conflicts of interest about its publication.

Declaration of Artificial Intelligence (AI) Assistance

The author used generative AI tools (e.g., ChatGPT 3.5) to improve the language and fluency of the manuscript. All AI-assisted content was thoroughly reviewed, revised, and confirmed by the author to ensure accuracy, coherence, and originality. The author takes full responsibility for the final publication.

Ethics Approval

Not applicable.

Funding

This study was not funded by any academic institution involved.

References

1. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. PLoS Computational Biology. 2006 Nov; 2(11):e155.
2. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Analyzing protein structure and function. InMolecular Biology of the Cell. Garland Science. 4th ed. 2002. https://www.ncbi.nlm.nih.gov/books/NBK26820/
3. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019 Nov; 20(11):681-97.
4. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function, and Bioinformatics. 1995 Apr; 21(4):319-344.
5. Yu ZG, Zhan XW, Han GS, Wang RW, Anh V, Chu KH. Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. International Journal of Molecular Sciences. 2010 Mar; 11(3):1141–1154
6. Dey S, Ghosh P, Das S. Positional difference and frequency (PdF) based alignment-free technique for genome sequence comparison. Journal of Biomolecular Structure and Dynamics. 2024 Dec; 42(23):12660-88.
7. Malod-Dognin N, Ban K, Pržulj N. Unified alignment of protein-protein interaction networks. Scientific Reports. 2017 Apr 19; 7(1):953.
8. Pal D, Dey S, Ghosh P, Bhattacharya DK, Das S, Maji B. A unique approach for protein secondary structure comparison under TOPS representation. Journal of Biomolecular Structure and Dynamics. 2024 Apr

22:1-3. https://doi.org/10.1080/07391102.2024.2333449
9. Seok C, Baek M, Steinegger M, Park H, Lee GR, Won J. Accurate protein structure prediction: what comes next. Biodesign. 2021 Sep; 9(3):47-50.
10. Bertoline LM, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: An overview of protein structure prediction. Frontiers in bioinformatics. 2023 Feb; 3:1120370.
11. AlQuraishi M. Machine learning in protein structure prediction. Current Opinion in Chemical Biology. 2021 Dec; 65:1-8.
12. Pakhrin SC, Shrestha B, Adhikari B, Kc DB. Deep learning-based advances in protein structure prediction. International Journal of Molecular Sciences. 2021. May 24; 22(11):5553.
13. Lyu Z, Wang Z, Luo F, Shuai J, Huang Y. Protein secondary structure prediction with a reductive deep learning method. Frontiers in Bioengineering and Biotechnology. 2021 Jun 15;9:687426.
14. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdritz G, Zhang J, Church GM, Sorger PK. Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology. 2022 Nov; 40(11):1617-23.
15. Yeung W, Zhou Z, Li S, Kannan N. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. Briefings in Bioinformatics. 2023 Jan; 24(1):bbac599.
16. Yang JY, Peng ZL, Chen X. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics. 2010 Jan 18; 11(Suppl 1):S9.
17. Abbasi NA, Akan OB. An information theoretical analysis of human insulin-glucose system toward the Internet of Bio-Nano Things. IEEE Transactions on Nanobioscience. 2017 Oct 11; 16(8):783–791.
18. Kha QH, Ho QT, Le NQ. Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles. Journal of Chemical Information and Modeling. 2022 Sep; 62(19):4820-6.
19. Xiao X, Wang P, Lin WZ, Jia JH, Chou KC. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical Biochemistry. 2013 May; 436(2):168-77.
20. Plötz T, Hammerla NY, Olivier P. Feature learning for activity recognition in ubiquitous computing. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2011 Jul; 22(1):1729-1734.
21. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Research. 1998 May; 26(9):2230-6.
22. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Science. 1997 Jan; 6(1):53-64.
23. Rahman MS, Shatabda S, Saha S, Kaykobad M, Rahman MS. DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC. Journal of Theoretical Biology. 2018 Sep 7; 452:22-34.
24. He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, Chou KC. Predicting drug-target interaction networks based on functional groups and biological features. PloS one. 2010 Mar 11;5(3):e9603.
25. Chou KC, Zhang CT. Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology. 1995; 30(4):275-349.
26. Levy A, Shalom BR, Chalamish M. A guide to similarity measures. arXiv preprint arXiv:2408.07706. 2024 Aug. https://arxiv.org/pdf/2408.07706
27. Roszkowska E, Filipowicz-Chomko M, Łyczkowska-Hanćkowiak A, Majewska E. Extended Hellwig's method utilizing entropy-based weights and mahalanobis distance: applications in evaluating sustainable development in the education area. Entropy. 2024 Feb 25;26(3):197.
28. Ellaway JI, Anyango S, Nair S, Zaki HA, Nadzirin N, Powell HR, Gutmanas A, Varadi M, Velankar S. Identifying protein conformational states in the Protein Data Bank: toward unlocking the potential of integrative dynamics studies. Structural Dynamics. 2024 May 1;11(3). https://pubs.aip.org/aca/sdy/article/11/3/034701/3294234
29. Hayashida M, Akutsu T. Measuring the similarity of protein structures using image compression algorithms. IEICE Transactions on Information and Systems. 2011 Dec; 94(12):2468-2478.
30. Thekkeveedu RP, Hegde S. Physicochemical properties and homology studies of the floral meristem identity gene LFY in nonflowering and flowering plants. BioTechnologia. 2022 Jun 29;103(2):113-129.

How to Cite: Saha S, Ghosh P, Ghosh D, Das S, Mondal MA, Singh D. Performance analysis of traditional distance metrics in protein structural class prediction. Int Res J Multidiscip Scope. 2025; 6(4):1103-1116. doi: 10.47857/irjms.2025.v06i04.06682