JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 01, Issue No. 02, 2025

ARTICLE TYPE: Research Article

Article ID: JEPCS-2025-0102-0025

iTransaminase-PseAAC: Identification of Transaminase and Non-Transaminase Sites in Proteins Using Position-Relative Features and Statistical Moments

Mahreen Zainab[1,*]

[1] School of Systems and Technology, University of Management and Technology, Lahore, Pakistan

*Corresponding author: f2018279036@umt.edu.pk

Contributing author: f2018279036@umt.edu.pk

Received: April 10, 2025 **Accepted:** June 21, 2025

Abstract- Accurately identifying transaminases from protein sequences helps in computational protein annotation. Transaminases are significant enzymes involved in amino-group transfer processes. This paper introduces iTransaminase-PseAAC, a sequence-based framework that uses statistical and position-relative properties to differentiate between transaminase and non-transaminase proteins. 3,077 transaminase and 2,500 non-transaminase protein sequences were included in the original dataset, which was gathered from UniProt. 1,002 positive and 2,129 negative sequences were obtained after using CD-HIT with a similarity threshold of 0.65 to remove sequence redundancy. Statistical moments were used in conjunction with the Position Relative Incidence Matrix (PRIM), Reverse Position Relative Incidence Matrix (RPRIM), Accumulative Absolute Position Incidence Vector (AAPIV), and Reverse Accumulative Absolute Position Incidence Vector (RAAPIV) to describe protein sequences. A Random Forest classifier was fed these characteristics, which capture amino acid composition, positional connections, and sequence orientation. Jackknife validation, 10-fold cross-validation, and self-consistency were used to assess the proposed model. For self-consistency testing and jackknife validation, the reported accuracies were 99.90% and 99.93%, respectively, with comparable excellent performance noted throughout the cross-validation folds. Additionally, a web interface for sequence-level prediction was created using Python. The results show that an efficient representation for transaminase classification on the reported benchmark may be obtained by combining statistical moments with position-relative sequence descriptors.

Keywords- Transaminase; Protein sequence classification; Random Forest; PseAAC; Statistical moments

1. Introduction

Sequence-based protein classification is a recurring problem in computational biology because experimentally determining protein function is costly and time-consuming. Pseudo-amino-acid-composition (PseAAC) representations and sequence-derived descriptors have therefore been used to encode information beyond simple residue frequencies [1]-[5]. The present work applies this general idea to the identification of transaminase proteins. Transaminases, also called aminotransferases, catalyze amino-group transfer between amino acids and keto acids. Their biochemical roles and substrate preferences have been studied experimentally and computationally [6]-[9]. Serum transaminase levels are also used clinically as biochemical indicators in several conditions, although the computational task addressed here is different: this paper predicts whether a *protein sequence* belongs to the transaminase class rather than diagnosing disease from patient measurements [10]-[13]. The problem is to derive a sequence representation that captures both composition and positional information, then use that representation for supervised classification. The proposed iTransaminase-PseAAC framework

combines statistical moments with PRIM, RPRIM, AAPIV, and RAAPIV features and trains a Random Forest classifier. The evaluation uses self-consistency, leave-one-out (jackknife), and 10-fold cross-validation.

This paper's primary contributions are:

- A sequence representation that combines statistical moments with forward and reverse position-relative features
- A Random Forest-based classifier evaluated using three validation procedures
- A Python-based web interface for sequence-level prediction
- A creation of a transaminase/non-transaminase benchmark from UniProt, followed by CD-HIT redundancy reduction at a 0.65 similarity cutoff.

The rest of the paper is organized as follows. The biological and computational background required to understand the sequence representation used in this investigation is given in Section 2. Related research on transaminases and sequence-based protein prediction is reviewed in Section 3. The Random Forest model, feature creation, and benchmark dataset are explained in Section 4. Section 5 presents the evaluation protocol. Section 6 reports and discusses the results, including the web interface and limitations. Section 7 concludes the paper.

2. Biological and Computational Background

The prediction problem addressed in this work starts from the primary structure of a protein. Proteins are chains of amino-acid residues arranged in a specific order, and this order provides the sequence information used by computational predictors. Proteins are functionally diverse macromolecules that perform catalytic, transport, structural, signaling, defense, contraction, and storage roles [14], [15]. For the present paper, the important point is not the full diversity of protein biology, but the fact that a protein can be represented as an ordered sequence whose composition and residue positions can be converted into numerical descriptors.

2.1. Amino Acids and Primary Protein Structure

The standard protein alphabet contains 20 common amino acids. Each amino acid has a central carbon atom attached to an amino group, a carboxyl group, a hydrogen atom, and a side chain, commonly denoted by R. The R group distinguishes one amino acid from another and contributes to chemical properties such as polarity, charge, and hydrophobicity [16], [17]. While Figure 2 summarizes the 20 amino acids based on the chemical composition of their side chains, the Figure 1 shows the common structural pattern of an amino acid.

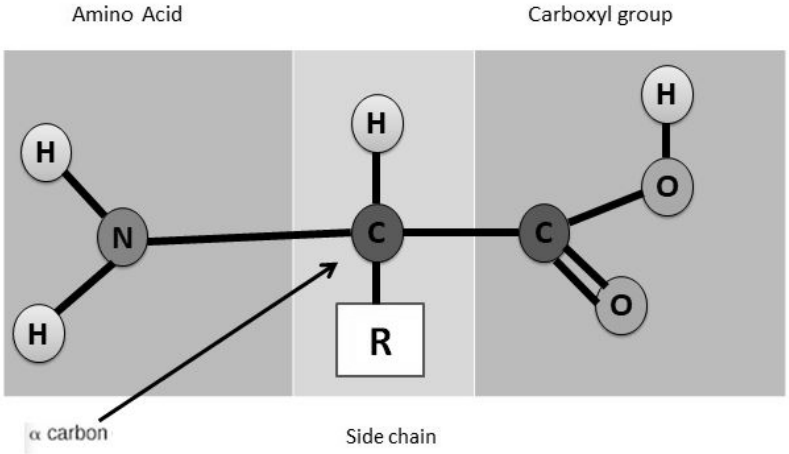


Figure 1: *General structural organization of an amino acid used as the basic unit of a protein sequence.*

The sequence is more than just a count of residues from a computational standpoint. Although the amino acid content of two proteins may be similar, the arrangement of those residues along the chain may differ. Feature representations that preserve both local or relative positional information and global composition are motivated by this difference. The purpose of the position-sensitive descriptors used later in this paper, namely PRIM, RPRIM, AAPIV, and RAAPIV, is to retain information that would be lost with a straightforward frequency vector.

2.2. Peptide Bonds, Directionality, and Sequence Order

Adjacent amino acids are connected through peptide bonds formed between the carboxyl group of one residue and the amino group of the next residue. This produces a polypeptide chain with a defined direction from the amino terminus (N-terminus) to the carboxyl terminus (C-terminus) [18]. The peptide-bond arrangement is illustrated in Figure 3.

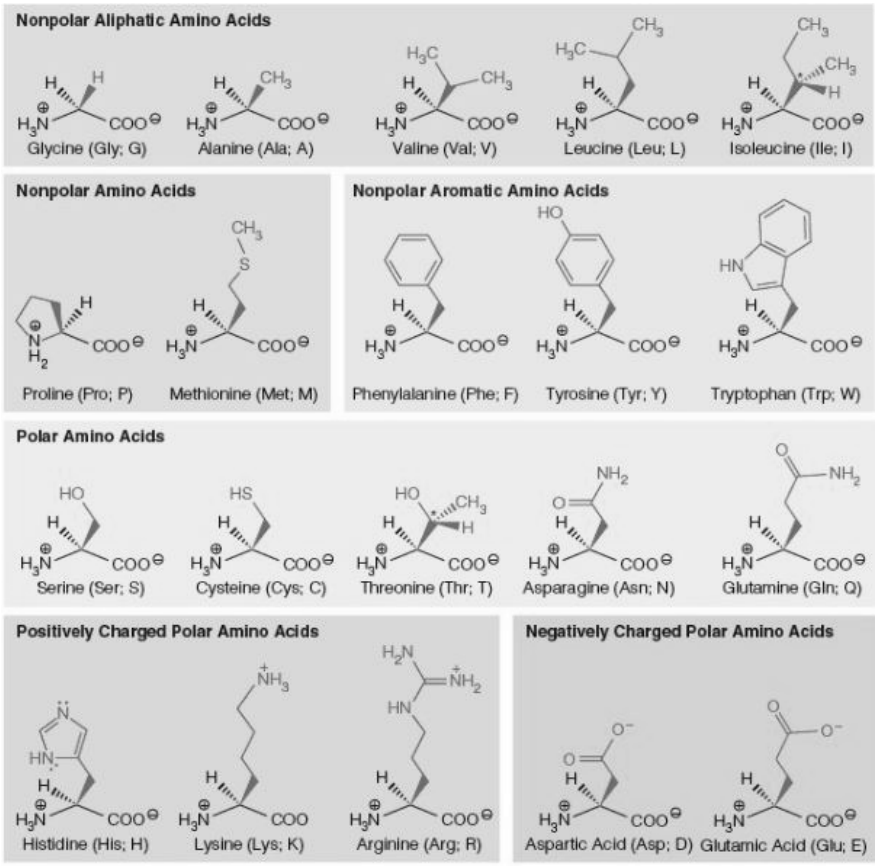


Figure 2: *The 20 common amino acids grouped according to side-chain properties.*

The directional nature of a protein sequence is relevant to the feature design used in this work. PRIM records relative residue positions in the original sequence, whereas RPRIM repeats the representation after reversing the sequence. Similarly, AAPIV and RAAPIV summarize accumulated positions in the forward and reverse directions. The use of both directions is therefore connected to an explicit property of the input representation: a protein sequence is ordered, and reversing that order changes the positional relationships among residues even when the amino-acid counts remain unchanged.

2.3. Transaminases and Amino-Group Transfer

Transaminases, or aminotransferases, catalyze transamination reactions in which an amino group is transferred between an amino acid and a keto acid [6], [7]. The reaction is reversible and forms an important part of amino-acid metabolism. A simplified transamination reaction is shown in Figure 4. Pyridoxal-5-phosphate (PLP) is an important cofactor in transaminase reactions. In the catalytic cycle, PLP participates in reversible Schiff-base formation and supports transfer of the amino group between donor and acceptor substrates [19]. Figure 5 illustrates the amino-acid-PLP Schiff-base intermediate involved in the catalytic cycle. Although they aren't used as direct inputs to the model, these biochemical features provide the categorization job context. The proposed predictor does not evaluate enzyme activity or mimic the catalytic process. Rather, it learns from protein sequences that have been classified as transaminase or non-transaminase and uses statistical and positional information acquired from the sequences to try to differentiate between the two groups.

2.4. Biomedical Context and Scope of the Predictor

Because blood alanine aminotransferase (ALT) and aspartate aminotransferase (AST) measures are utilized to evaluate liver and other metabolic disorders, transaminases are also extensively studied in biomedical literature [10]–[13], [20]. Abnormal transaminase levels have been linked to insulin resistance, non-alcoholic fatty liver disease, and other diseases in clinical investigations. The biological significance of transaminases is explained by that clinical literature, but it should not be confused with this paper's computational goal. Amino acid sequences from a public protein database are used by the predictor presented here. Its output is a sequence-level classification between transaminase and non-transaminase proteins. It does not take serum ALT or AST measurements as input, does not estimate disease risk, and is not evaluated as a clinical diagnostic system. Maintaining this distinction is important when interpreting the results and the potential

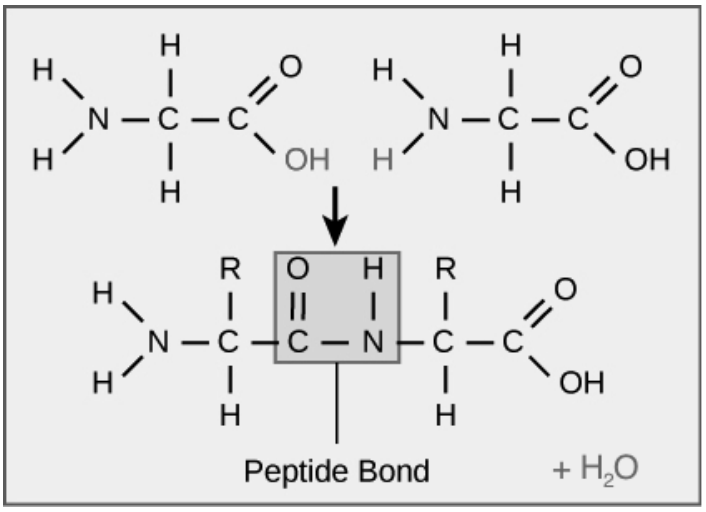


Figure 3: *Formation of a peptide bond between two amino acids.*

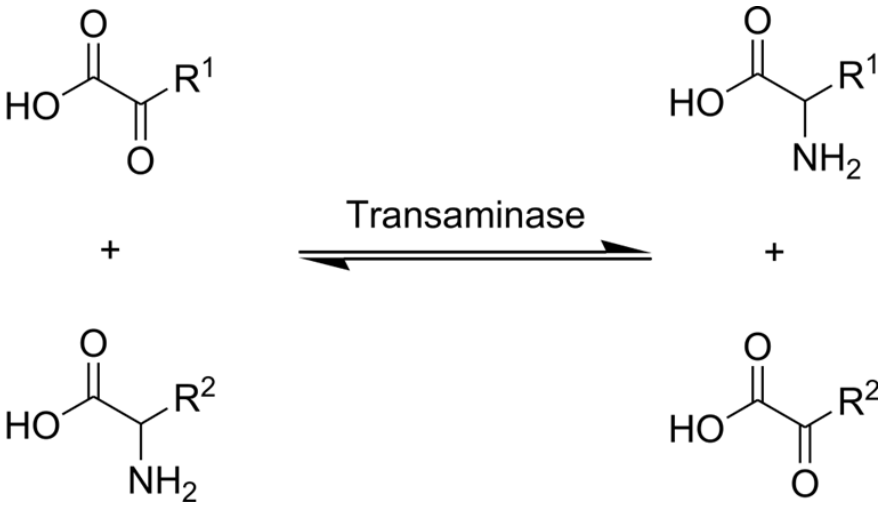


Figure 4: *General transamination reaction between an amino acid and an α-keto acid.*

use of the web interface. Table 1 summarizes how the biological background connects to the computational representation used by iTransaminase-PseAAC.

Table 1: *Connection between biological information and the sequence-based representation used in this paper.*

Biological aspect	Information available from sequence	Representation in the model
Amino-acid composition	Residue identities and occurrence frequencies	Frequency-derived information and statistical moments
Sequence order	Relative location of residues along the polypeptide chain	PRIM and AAPIV
Sequence direction	Positional relationships viewed from the opposite sequence direction	RPRIM and RAAPIV
Protein-level pattern	Combined composition and position information across the complete sequence	Integrated feature vector supplied to the Random Forest
Transaminase class label	Annotation of a sequence as transaminase or non-transaminase	Supervised target class

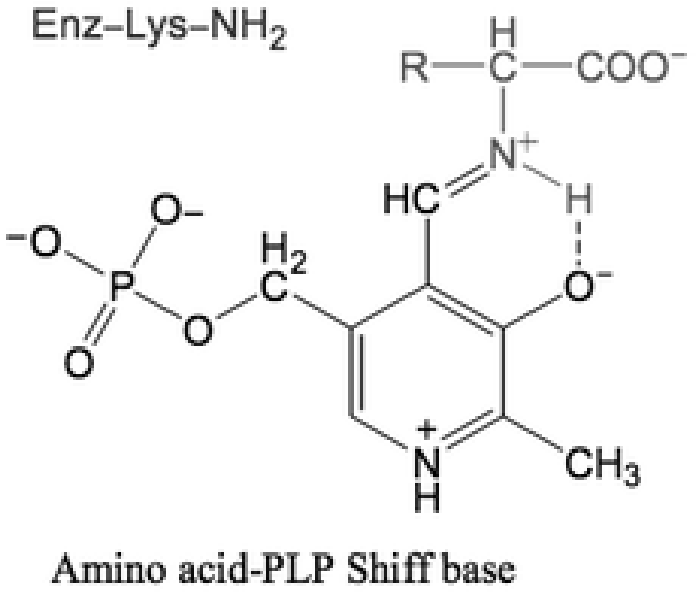


Figure 5: *Amino-acid-PLP Schiff-base intermediate associated with transaminase catalysis.*

3. Related Work

Recent transaminase studies have examined enzyme function, substrate scope, engineering, and biomedical relevance. Voss et al. investigated the emergence of (*R*)-amine transaminase activity within an amino-acid-transaminase scaffold, while Tang et al. characterized a transaminase from *Rhizobium* sp. [6], [7]. Cheng et al. used high-throughput screening for transaminase engineering and substrate-scope extension [8]. Bowsher and Henry characterized rat-brain cytosolic tyrosine transaminase as glutamine transaminase-K [9]. Other studies considered transaminase activity in clinical or biochemical settings. Zhou et al. examined relationships among folate, vitamin B12, homocysteine, transaminases, and mild cognitive impairment, while Brumboiu et al. analyzed pathological serum-transaminase levels [10], [11]. Dai et al. reviewed acupoint embedding for non-alcoholic fatty liver disease with abnormal transaminase, and Yamamoto et al. studied the association between serum transaminases and insulin resistance [12], [13]. Transaminases have also been studied as biocatalysts and therapeutic targets. Matassa et al. reported a transaminase-catalyzed synthesis and recovery process for a chiral amine, Meng et al. investigated computational stabilization of ω-transaminases, and Matsumoto et al. developed a transaminase-mediated pathway for n-butylamine production [21]-[23]. Yoshida et al. investigated inhibition of glutamate-oxaloacetate transaminase 1 in pancreatic cancer, while Planchestainer et al. described a screening strategy for transaminase evolution [24], [25]. Hegde et al. reported GABA transaminase deficiency with survival into adulthood [26]. For sequence-based protein prediction, PseAAC-style representations and position-sensitive descriptors have been used in several protein-property prediction tasks. Examples include N-linked glycosylation prediction, S-palmitoylation prediction, disulfide-bonding prediction, phosphotyrosine prediction, and protease classification [2]-[5], [27]. These studies motivate the use of sequence-derived statistical and positional features in the present classifier. A common pattern across these sequence-based predictors is the conversion of a variable-length amino-acid sequence into a fixed-length numerical representation before classification. While position-relative features preserve information about residue locations, statistical moments offer global descriptors of the mapped sequence. For protein structure or functional prediction, related research have integrated PseAAC-style characteristics with ensemble approaches, support vector machines, and other classifiers [28]–[33]. These papers give the current feature-construction technique methodological background. Additionally, the literature distinguishes between two directions of research. Biochemical characterisation, enzyme engineering, catalytic activity, and clinical connections are major topics in the transaminase literature [6]–[26]. On the other hand, computer prediction of protein characteristics from sequence is the subject of PseAAC-based investigations [1]–[5], [27], [28]–[33]. By building a transaminase-specific sequence classifier that integrates position-relative descriptors, statistical moments, and a Random Forest model, the current study links these directions. Therefore, rather than measuring transaminase activity in a lab, the paper focuses on protein annotation from sequence.

4. Materials and Methods

4.1. Benchmark Dataset

Transaminase-related annotations were used to gather protein sequences from the UniProt database. Prior to redundancy reduction, the original collection comprised 2,500 non-transaminase sequences and 3,077 transaminase sequences. Sequences from organisms like Arabidopsis thaliana, Bacillus subtilis, human, Escherichia coli K12, and mouse were included in the positive collection, whereas sequences from A. thaliana, human, mouse, rat, and Saccharomyces cerevisiae were included in the non-transaminase collection. Prior to model creation, entries with ambiguous or incomplete annotations were eliminated.

When similarly related proteins are included in both model-development and evaluation subsets, redundant and extremely similar sequences may inflate performance estimates. To reduce homology bias, CD-HIT was used [34]. There is a claimed similarity threshold of 0.65. Following this stage, there were 2,129 negative and 1,002 positive sequences in the sample. The reduced dataset is the more appropriate benchmark for reporting model generalization, although the experimental description does not state with complete clarity which reported validation tables were produced before or after this filtering step. That point is retained as a reproducibility issue in Section 6.6.

Let the benchmark set be written as

$$\mathcal{S} = \mathcal{S}^{+} \cup \mathcal{S}^{-}, \tag{1}$$

where $\mathcal{S}^{+}$ denotes transaminase sequences and $\mathcal{S}^{-}$ denotes non-transaminase sequences.

The overall prediction workflow is shown in Figure 6.

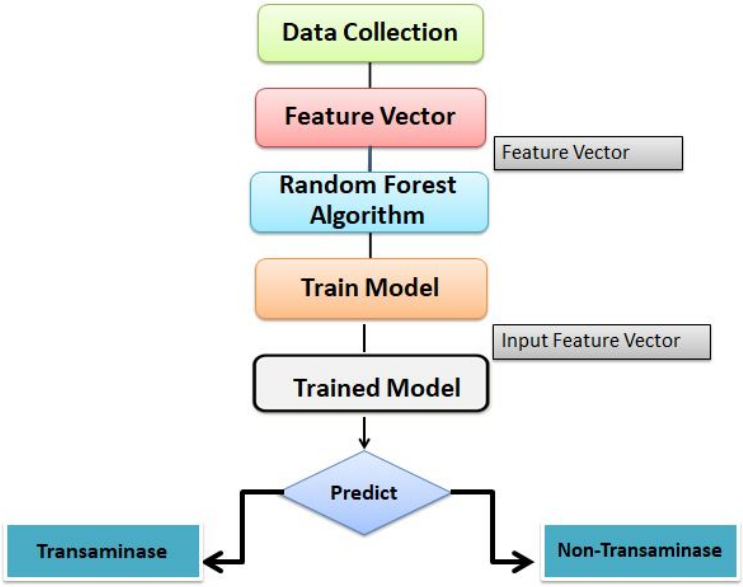


Figure 6: *Workflow used to construct and evaluate the transaminase prediction model.*

4.2. Sequence Representation

The feature design follows the general sequence-representation strategy used in PseAAC-related predictors [1], [35]. The source sequence is mapped into numerical form so that composition and positional information can be represented within a fixed-length feature vector. If a protein sequence of length m is written as

$$S = \{\beta_1, \beta_2, \ldots, \beta_m\}, \tag{2}$$

then numerical values are assigned to residues and the one-dimensional sequence is mapped into a two-dimensional arrangement. The representation uses

$$z = \lceil \sqrt{m} \rceil \tag{3}$$

as the dimensioning term for this mapping. The two-dimensional representation is then used to calculate moments, while separate position-relative descriptors are derived directly from the residue order. This combination allows the final vector to include global distributional information together with explicit sequence-position information.

The paper combines five feature components:

1. statistical moments;
2. Position Relative Incidence Matrix (PRIM);
3. Reverse Position Relative Incidence Matrix (RPRIM);
4. Accumulative Absolute Position Incidence Vector (AAPIV); and
5. Reverse Accumulative Absolute Position Incidence Vector (RAAPIV).

Related work on high-dimensional and sequence-based learning provides context for the use of compact numerical descriptors [36]-[38]. Table 2 summarizes the role of each feature family in the final representation.

Table 2: *Feature components used to represent each protein sequence.*

Feature component	Source information	Purpose in the representation
Raw moments	Two-dimensional numerical mapping of the sequence	Summarize distributional properties of the mapped values
Central moments	Sequence values measured around the centroid	Describe variation relative to the center of the mapped representation
Hahn moments	Orthogonal polynomial representation	Add order-sensitive descriptors of the mapped sequence
PRIM	Relative positions in the forward sequence	Encode pairwise positional incidence
RPRIM	Relative positions in the reversed sequence	Encode positional incidence from the reverse direction
AAPIV	Accumulated absolute residue positions	Retain forward positional information in vector form
RAAPIV	Accumulated positions in the reversed sequence	Retain reverse positional information in vector form

4.3. Statistical Moments

The representation uses raw, central, and Hahn moments. A protein sequence is first mapped to a two-dimensional numerical representation. For a mapped matrix a_{xy}, the raw moment of order (p, q) is represented as

$$M_{pq} = \sum_{x} \sum_{y} x^p y^q a_{xy}. \tag{4}$$

The centroid is obtained from the first-order raw moments,

$$\bar{x} = \frac{M_{10}}{M_{00}}, \qquad \bar{y} = \frac{M_{01}}{M_{00}}, \tag{5}$$

and the corresponding central moment is

$$\mu_{pq} = \sum_{x} \sum_{y} (x - \bar{x})^p (y - \bar{y})^q a_{xy}. \tag{6}$$

Moments are calculated up to third order. For the raw and central families, this produces a compact set of descriptors such as M_{00}, M_{01}, M_{10}, and higher-order combinations whose total order does not exceed three. The raw moments describe the distribution of the mapped values, while the central moments are evaluated relative to the centroid and therefore emphasize variation around the center of the numerical representation. Hahn moments are also used to obtain additional order-sensitive descriptors. Hahn moments are derived from discrete orthogonal polynomials and are used here as another compact description of the two-dimensional sequence mapping. Similar moment-based representations have been used in earlier pattern-recognition and protein-prediction studies [27], [39], [40]. The three moment families are not treated as independent classifiers; their coefficients are concatenated with the position-relative features to form the representation supplied to the Random Forest.

4.4. Position-Relative Features

PRIM captures relative positional information among the 20 standard amino acids. The resulting matrix has dimension 20×20, so each row and column corresponds to one amino-acid type. An entry summarizes the relative incidence of

one residue type with respect to another in the original sequence. Because the complete matrix contains 400 coefficients, moment calculations are used to obtain a more compact representation.

RPRIM applies the same construction after reversing the primary sequence. This does not change the amino-acid composition, but it changes the order in which positional relationships are encountered. Using both PRIM and RPRIM therefore provides complementary views of the same sequence without requiring an explicit structural model. The reduced positional descriptors are then combined with the other feature families described in Table 2.

The frequency vector records residue occurrence counts,

$$\boldsymbol{\xi} = \{\tau_1, \tau_2, \ldots, \tau_{20}\}, \tag{7}$$

where τ_i is the frequency of the ith amino acid. AAPIV records accumulated residue positions in the original sequence. For residue type i occurring at positions $r_1, \ldots, r_n$, its component is represented as

$$v_i = \sum_{t=1}^{n} r_t. \tag{8}$$

RAAPIV applies the same construction to the reversed sequence. If a residue occurs repeatedly at different locations, AAPIV and RAAPIV preserve the accumulated positional contribution instead of reducing the sequence to a simple count. Position-relative descriptors of this type are intended to retain information that is lost when only global residue frequencies are used [2], [3], [41]. The final feature vector is constructed by integrating the moment coefficients with PRIM/RPRIM-derived descriptors, residue-frequency information, and the AAPIV/RAAPIV components. The final dimensionality of the concatenated vector is not reported and should be documented for reproducibility. The important design choice is that the representation combines composition, relative position, and sequence direction before classification.

4.5. Random Forest Classifier

The extracted feature vectors are classified using a Random Forest. A Random Forest combines multiple decision trees and assigns the final class by aggregating the individual tree outputs. Figure 7 shows the Random Forest ensemble structure used for classification.

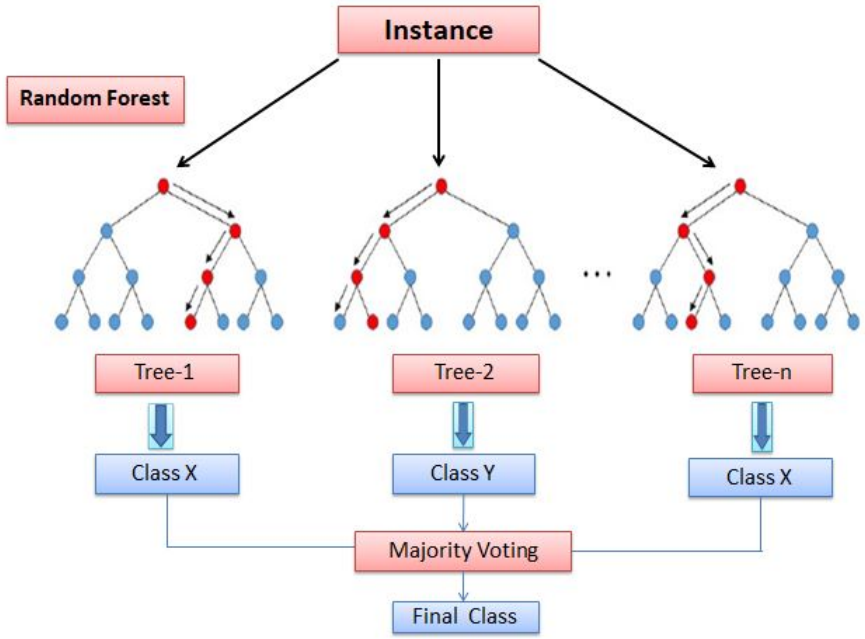


Figure 7: *Random Forest ensemble used for classification.*

Classification challenges involving diverse predictors and high-dimensional feature sets have been addressed with Random Forests [42]–[44]. Another example of this classifier in computational biology is protein-binding-site prediction using Random Forest [45]. Samples of training data are used to generate several decision trees. Instead of choosing candidate features from the whole feature set, a random subset is chosen at each stage of the tree construction process. The forest uses majority voting to aggregate the class judgments made by each tree. Because the input vector includes

Table 3: *Self-consistency results for iTransaminase-PseAAC.*

Predictor	Acc (%)	Sp (%)	Sn (%)	MCC
iTransaminase-PseAAC	99.9	100	99.9	0.9985

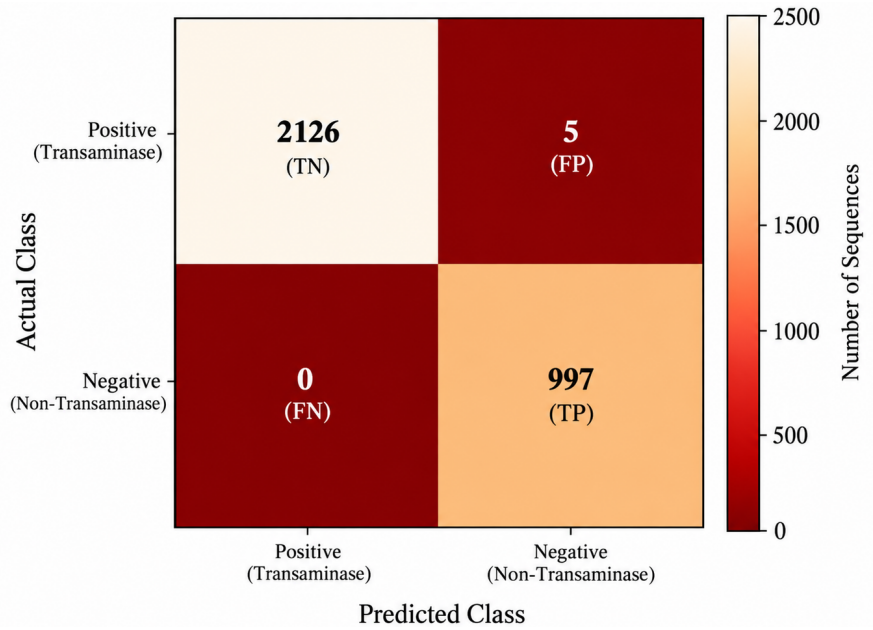


Figure 8: *Confusion matrix reported for self-consistency testing.*

many feature families with distinct numerical properties, this ensemble structure is helpful for the current representation. Four main processes form the training process: choose training observations, build a decision tree, repeat the procedure to get the necessary number of trees, and combine the tree outputs to create a new feature vector. The present experimental description does not include the precise number of trees, tree-depth restrictions, random seed, or feature-subsampling parameters. Because they are necessary for precise repeatability, these implementation details should be extracted from the source code prior to final publishing.

5. Experimental Evaluation

5.1. Evaluation Metrics

The paper reports sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC). Using the standard confusion-matrix notation TP, TN, FP, and FN, these metrics are

$$Sn = \frac{TP}{TP + FN}, \tag{9}$$

$$Sp = \frac{TN}{TN + FP}, \tag{10}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{12}$$

These metrics are commonly used in sequence-based prediction studies to assess positive-class recovery, negative-class recovery, overall accuracy, and balanced correlation [46]-[48].

5.2. Validation Protocols

Three procedures are reported.

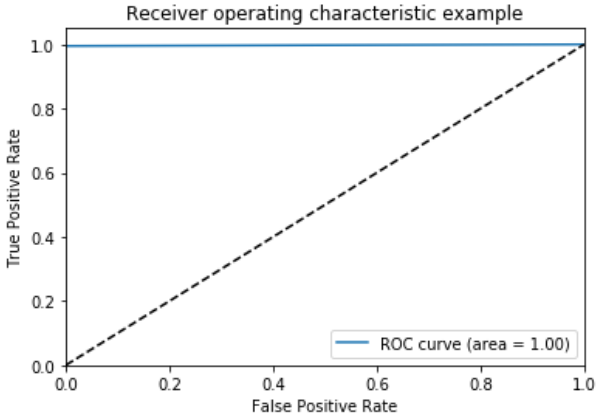


Figure 9: *ROC curve reported for self-consistency testing.*

Table 4: *Jackknife validation results for iTransaminase-PseAAC.*

Predictor	Acc (%)	Sp (%)	Sn (%)	MCC
iTransaminase-PseAAC	99.93	99.90	99.95	0.52

Self-consistency testing evaluates the model on the same data used for model construction. It measures how well the fitted model reproduces its training labels, but it does not provide an independent estimate of generalization.

Jackknife validation follows a leave-one-out procedure in which the model is trained on all but one sample and tested on the held-out sample. The process is repeated across the benchmark [28], [29], [49].

10-fold cross-validation partitions the benchmark into ten folds, trains on nine folds, and evaluates on the remaining fold. The procedure is repeated so that each fold is used for evaluation. Cross-validation and related resampling strategies are widely used in protein-sequence prediction [30]-[33].

5.3. Interpretation of the Validation Procedures

The three reported procedures answer different questions and should therefore not be interpreted as interchangeable evidence. Self-consistency testing measures whether the fitted classifier can reproduce the labels of the data used during model construction. It is useful as a training check but is expected to be optimistic because the same observations are used for fitting and evaluation. Jackknife validation is more stringent because each sequence is held out once, while 10-fold cross-validation evaluates groups of held-out sequences and is less computationally demanding than leave-one-out validation. The benchmark is also imbalanced after CD-HIT filtering, with 1,002 reported positive sequences and 2,129 reported negative sequences. Accuracy alone therefore does not provide a complete description of classifier behavior. Sensitivity and specificity separate performance on the two classes, while MCC summarizes the relationship between predicted and observed labels in a single value. For this reason, the discussion below considers the complete set of reported metrics rather than relying only on accuracy.

6. Results and Discussion

Table 3 reports the self-consistency results and the jackknife results are reported in Table 4. The corresponding confusion matrix is shown in Figure 8, and the source ROC plot is shown in Figure 9. Because self-consistency reuses the training data for evaluation, the near-perfect result should be treated as a fit check rather than an estimate of performance on unseen sequences. Table 5 reports the fold-wise cross-validation values. Table 6 compares the proposed approach with other approaches.

6.1. Web Interface

A Python-based web interface was implemented for entering a protein sequence and obtaining a transaminase/non-transaminase prediction. Related PseAAC studies have emphasized accessible web implementations for computational predictors [50]-[53]. The interface represents the deployment stage of the workflow: a user supplies an amino-acid sequence in the prediction field, submits the sequence, and receives the predicted class from the trained model. A screenshot of the reported interface is shown in Figure 10. Separate sections for general information, prediction, work-related information, and other data are included in the interface.

Table 5: *Ten-fold cross-validation results.*

Fold	Sn (%)	Sp (%)	MCC	Acc (%)
K1	100	100	1.0000	100
K2	100	100	1.0000	100
K3	99	100	0.9927	99.7
K4	100	100	1.0000	100
K5	100	100	1.0000	100
K6	100	100	1.0000	100
K7	99	100	0.9927	99.7
K8	100	100	1.0000	100
K9	99	100	0.9926	99.7
K10	100	100	1.0000	100
Average	100	99.77	0.9978	99.91

Table 6: *Comparison with other methods.*

Method	Sensitivity	Specificity	Accuracy	MCC
SVM BUPA LIVER	80%	37.5%	63%	0.37
SVM ILPD	95%	10.8%	71%	0.21
MLP	100%	100%	100%	1.0
GRNN	100%	98.99%	99.50%	1.00
NeuroSVM	-	-	99.83%	-

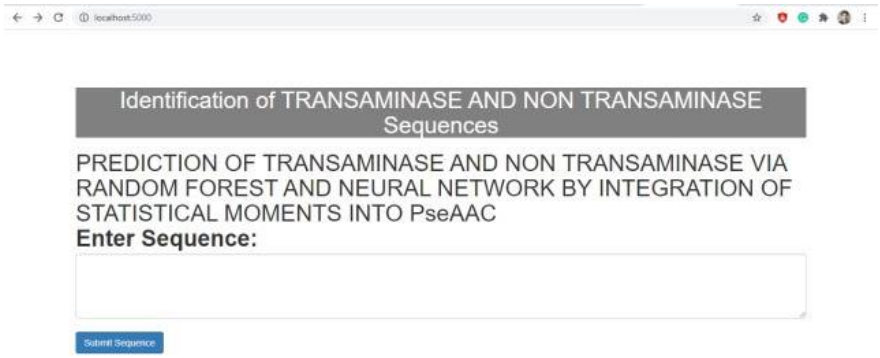


Figure 10: *Web interface reported for iTransaminase-PseAAC.*

6.2. Limitations

Although the results are satisfactory, there are a few limitations. First, the precise samples utilized in each validation experiment should be made clear, as both pre-CD-HIT and post-CD-HIT dataset sizes are presented. Second, there is a discrepancy between the independent cross-validation summary and the 10-fold average. Third, the reported jackknife MCC of 0.52 should be compared to the original output as it seems abnormally low in comparison to the claimed accuracy, sensitivity, and specificity. Lastly, it is unclear from the comparison table if the techniques given were separated and evaluated on the same dataset.

7. Conclusion and Future Work

iTransaminase-PseAAC, a sequence-based computational framework for differentiating between transaminase and non-transaminase proteins, has been proposed in this paper. In order to express amino acid composition, relative residue locations, and sequence orientation inside the feature vector provided to a Random Forest classifier, the suggested representation integrates statistical moments with PRIM, RPRIM, AAPIV, and RAAPIV descriptors. A

reduced benchmark with 1,002 positive and 2,129 negative sequences was created via CD-HIT redundancy reduction at a similarity threshold of 0.65 from the original sample, which had 3,077 transaminase and 2,500 non-transaminase sequences. Self-consistency, jackknife, and 10-fold cross-validation were used to assess the model. The fold-wise 10-fold cross-validation results gave an average accuracy of 99.91%, whereas the reported accuracies for self-consistency testing and jackknife validation were 99.90% and 99.93%, respectively. These results collectively show that the two protein classes on the presented benchmark may be successfully distinguished using the combined statistical and position-relative representation. The trained predictor was also made available for sequence-level categorization through the implementation of a Python-based web interface. Future work should therefore emphasize reproducible model settings, clearly documented feature dimensionality, and evaluation on an independent non-redundant protein benchmark. Such evaluation would provide stronger evidence of how well iTransaminase-PseAAC generalizes beyond the sequences used in the present paper.

Author Declarations

Author Contributions (CRediT): Mahreen Zainab: Conceptualization, Methodology, Investigation, Literature Review, Data Curation, Formal Analysis, Visualization, Writing - Original Draft. The author reviewed and approved the final manuscript.

Funding: No external funding was received for this research.

Data Availability Statement: Protein sequences were collected from UniProt and supplementary benchmark data were used.

Acknowledgments: No acknowledgment information was provided by the author.

Conflicts of Interest: The author declare no conflicts of interest.

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this work, the author used generative AI tools only for language editing and proofreading, where applicable. The authors reviewed and edited the content and take full responsibility for the published article.

References

[1] Chou, Kuo Chen. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology, 273(1), 236–247. https://doi.org/10.1016/j.jtbi.2010.12.024

[2] Akmal, M. A., Rasool, N., & Khan, Y. D. (2017). Prediction of N-linked glycosylation sites using position relative features and statistical moments. PLoS ONE, 12(8), 1–21. https://doi.org/10.1371/journal.pone.0181966

[3] Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., & Chou, K. C. (2019). SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Analytical Biochemistry, 568(November 2018), 14–23. https://doi.org/10.1016/j.ab.2018.12.019

[4] Khan, Y. D., Jamil, M., Hussain, W., Rasool, N., Khan, S. A., & Chou, K. C. (2019). pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. Journal of Theoretical Biology, 463, 47–55. https://doi.org/10.1016/j.jtbi.2018.12.015

[5] Khan, Y. D., Amin, N., Hussain, W., Rasool, N., Khan, S. A., & Chou, K. C. (2020). iProtease-PseAAC(2L): A two-layer predictor for identifying proteases and their types using Chou's 5-step-rule and general PseAAC. Analytical Biochemistry, 588(February 2019), 113477. https://doi.org/10.1016/j.ab.2019.113477

[6] Tang, K., Yi, Y., Gao, Z., Jia, H., Li, Y., Cao, F., ... Wei, P. (2020). Identification , Heterologous Expression and Characterization of a Transaminase from Rhizobium sp . Catalysis Letters, (0123456789). https://doi.org/10.1007/s10562-020-03121-2

[7] Voß, M., Xiang, C., Esque, J., Nobili, A., Marian, J., André, I., ... Bornscheuer, U. T. (2020). within an # -amino acid transaminase scaffold Creation of (R) -amine transaminase activity within an α-amino acid transaminase scaffold. https://doi.org/10.1021/acschembio.9b00888

[8] Cheng, F., Chen, X. L., Xiang, C., Liu, Z. Q., Wang, Y. J., & Zheng, Y. G. (2020). Fluorescence-based high-throughput screening system for R-ω-transaminase engineering and its substrate scope extension. Applied Microbiology and Biotechnology, 104(7), 2999–3009. https://doi.org/10.1007/s00253-020-10444-y

[9] Bowsher, R. R., & Henry, D. P. (2020). Neurochemistry International Purification , characterization and identification of rat brain cytosolic tyrosine transaminase as glutamine Transaminase-K. Neurochemistry International, 133(August 2019), 104653. https://doi.org/10.1016/j.neuint.2019.104653

[10] Zhou, X., Wang, Q., An, P., Du, Y., Zhao, J., & Song, A. (2019). Relationship between folate , vitamin B 12 , homocysteine , transaminase and mild cognitive impairment in China : a case-control study. International Journal of Food Sciences and Nutrition, 0(0), 1–10. https://doi.org/10.1080/09637486.2019.1648387

[11] Brumboiu, M. I., Brice, P., Ndemba, A., & Cazacu, I. (2019). Computing tools for analysing the pathological levels of serum transaminases, 41(1), 24–32.

[12] Dai, L., Ooi, V. V., Zhou, W., Ji, G., & Abd-Elsalam, S. (2020). Acupoint embedding therapy improves nonalcoholic fatty liver disease with abnormal transaminase: A PRISMA-compliant systematic review and meta-analysis. Medicine (United States), 99(3). https://doi.org/10.1097/MD.0000000000018775

[13] Yamamoto, J. M., Padro-Nuñez, S., Guarnizo-Poma, M., Lazaro-Alcantara, H., Paico-Palacios, S., Pantoja-Torres, B., ... Benites-Zapata, V. A. (2020). Association between serum transaminase levels and insulin resistance in euthyroid and non-diabetic adults: Serum transaminase levels and insulin resistance in healthy adults. Diabetes and Metabolic Syndrome: Clinical Research and Reviews, 14(1), 17–21. https://doi.org/10.1016/j.dsx.2019.11.013

[14] F. H. Crick. (1958). On protein synthesis. Symposia of the Society for Experimental Biology, 12, 138–163. https://doi.org/10.1038/227561a0

[15] Finkelstein, M., & Weissmann, G. (1978). The introduction of enzymes into cells by means of liposomes. Journal of Lipid Research, 19(3), 289–303.

[16] khanacademy. (2020). proteins-and-amino-acids/v/introduction-to-amino-acids.

[17] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

[18] Sheehan, J. C., & Hess, G. P. (1955). A new method of forming peptide bonds. Journal of the American Chemical Society, 77(4), 1067–1068. https://doi.org/10.1021/ja01609a099

[19] VELICK, S. F., & VAVRA, J. (1962). A kinetic and equilibrium analysis of the glutamic oxaloacetate transaminase mechanism. The Journal of Biological Chemistry, 237(7), 2109–2122.

[20] Oh, R. C., Hustead, T. R., Army, T., Medicine, F., & Program, R. (2011). Causes and Evaluation of Mildly Elevated Liver Transaminase Levels, 1003–1008.

[21] Matassa, C., Ormerod, D., Bornscheuer, U. T., Höhne, M., & Satyawali, Y. (2020). Three-liquid-phase Spinning Reactor for the Transaminase-catalyzed Synthesis and Recovery of a Chiral Amine. ChemCatChem, 12(5), 1288–1291. https://doi.org/10.1002/cctc.201902056

[22] Meng, Q., Capra, N., Palacio, C. M., Lanfranchi, E., Otzen, M., Van Schie, L. Z., ... Janssen, D. B. (2020). Robust ω-Transaminases by Computational Stabilization of the Subunit Interface. ACS Catalysis, 10(5), 2915–2928. https://doi.org/10.1021/acscatal.9b05223

[23] Matsumoto, T., Mori, Y., Tanaka, T., & Kondo, A. (2020). n-Butylamine production from glucose using a transaminase-mediated synthetic pathway in Escherichia coli. Journal of Bioscience and Bioengineering, 129(1), 99–103. https://doi.org/10.1016/j.jbiosc.2019.06.015

[24] Yoshida, T., Yamasaki, S., Kaneko, O., Taoka, N., Tomimoto, Y., Namatame, I., ... Lyssiotis, C. A. (2020). A covalent small molecule inhibitor of glutamate-oxaloacetate transaminase 1 impairs pancreatic cancer growth. Biochemical and Biophysical Research Communications, 522(3), 633–638. https://doi.org/10.1016/j.bbrc.2019.11.130

[25] Planchestainer, M., Hegarty, E., Gourlay, L. J., Paradisi, F., & Heckmann, C. M. (2019). Chemical Science Widely applicable background depletion step enables transaminase evolution through solid- phase screening †, 5952–5958. https://doi.org/10.1039/c8sc05712e

[26] Hegde, A. U., Karnavat, P. K., Vyas, R., Dibacco, M. L., Grant, P. E., & Pearl, P. L. (2019). GABA Transaminase Deficiency With Survival Into Adulthood. https://doi.org/10.1177/0883073818823359

[27] Khan, Y. D., Rasool, N., Hussain, W., Khan, S. A., & Chou, K. C. (2018). iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. Molecular Biology Reports, 45(6), 2501–2509. https://doi.org/10.1007/s11033-018-4417-z

[28] Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. Journal of Theoretical Biology, 364, 284–294. https://doi.org/10.1016/j.jtbi.2014.09.029

[29] Dou, Y., Yao, B., & Zhang, C. (2014). PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. Amino Acids, 46(6), 1459–1469. https://doi.org/10.1007/s00726-014-1711-5

[30] Feng, K. Y., Cai, Y. D., & Chou, K. C. (2005). Boosting classifier for predicting protein domain structural class. Biochemical and Biophysical Research Communications, 334(1), 213–217. https://doi.org/10.1016/j.bbrc.2005.06.075

[31] Kumar, R., Srivastava, A., Kumari, B., & Kumar, M. (2015). Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology, 365, 96–103. https://doi.org/10.1016/j.jtbi.2014.10.008

[32] Mondal, S., & Pai, P. P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. Journal of Theoretical Biology, 356, 30–35. https://doi.org/10.1016/j.jtbi.2014.04.006

[33] Nanni, L., Brahnam, S., & Lumini, A. (2014). Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. Journal of Theoretical Biology, 360, 109–116. https://doi.org/10.1016/j.jtbi.2014.07.003

[34] Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

[35] Chou, K. C. (2001b). Using subsite coupling to predict signal peptides. Protein Engineering, 14(2), 75–79. https://doi.org/10.1093/protein/14.2.75

[36] Altay, G., & Emmert-Streib, F. (2010). Revealing differences in gene network inference algorithms on the network level by ensemble methods. Bioinformatics, 26(14), 1738–1744. https://doi.org/10.1093/bioinformatics/btq259

[37] Wan, S., Mak, M. W., & Kung, S. Y. (2016). Ensemble Linear Neighborhood Propagation for Predicting Subchloroplast Localization of Multi-Location Proteins. Journal of Proteome Research, 15(12), 4755–4762. https://doi.org/10.1021/acs.jproteome.6b00686

[38] Wan, S., Mak, M. W., & Kung, S. Y. (2017). Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(1), 212–224. https://doi.org/10.1109/TCBB.2016.2527657

[39] Khan, Y. D., Ahmad, F., & Anwar, M. W. (2012). A neuro-cognitive approach for Iris recognition using back propagation. World Applied Sciences Journal, 16(5), 678–685.

[40] Khan, Y. D., Khan, S. A., Ahmad, F., & Islam, S. (2014). Iris recognition using image moments and k-Means algorithm. The Scientific World Journal, 2014. https://doi.org/10.1155/2014/723595

[41] Wu, C. H. (1997). Artificial neural networks for molecular sequence analysis. Computers and Chemistry, 21(4), 237–256. https://doi.org/10.1016/S0097-8485(96)00038-1

[42] Bouaziz, A., Dartigues-Pallez, C., Da Costa Pereira, C., Precioso, F., & Lloret, P. (2014). Short text classification using semantic random forest. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8646 LNCS, 288–299. https://doi.org/10.1007/978-3-319-10160-6_26

[43] Pavlov, Y. L. (2019). Random forests. Random Forests, 1–122. https://doi.org/10.1201/9780367816377-11

[44] Bienvenido-huertas, D., Rubio-bellido, C., & P, J. L. (2020). Automation and optimization of in-situ assessment of wall thermal transmittance using a Random Forest algorithm, 168(August 2019). https://doi.org/10.1016/j.buildenv.2019.106479

[45] Liu, Z. P., Wu, L. Y., Wang, Y., Zhang, X. S., & Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics, 26(13), 1616–1622. https://doi.org/10.1093/bioinformatics/btq253

[46] Chou, K. C. (2001a). Prediction of signal peptides using scaled window. Peptides, 22(12), 1973–1979. https://doi.org/10.1016/S0196-9781(01)00540-X

[47] Feng, P., Ding, H., Chen, W., & Lin, H. (2013). Na ve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins, 2013.

[48] Xu, Y., Shao, X., Wu, L., & Deng, N. (2013). iSNO-AAPair : incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S -nitrosylation sites in proteins, 1–18. https://doi.org/10.7717/peerj.171

[49] Cai, Y. D., & Zhou, G. P. (2000). Prediction of protein structural classes by neural network. Biochimie, 82(8), 783–785. https://doi.org/10.1016/S0300-9084(00)01161-5

[50] Qiu, W. R., Jiang, S. Y., Xu, Z. C., Xiao, X., & Chou, K. C. (2017). iRNAm5C-PseDNC: Identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget, 8(25), 41178–41188. https://doi.org/10.18632/oncotarget.17104

[51] Qiu, W. R., Sun, B. Q., Xiao, X., Xu, D., & Chou, K. C. (2017). iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory. Molecular Informatics, 36(5), 1–10. https://doi.org/10.1002/minf.201600010

[52] Chou, K.-C. (2015). Impacts of Bioinformatics to Medicinal Chemistry. Medicinal Chemistry, 11(3), 218–234. https://doi.org/10.2174/1573406411666141229162834

[53] Sigma, S. (2005). N Atural S P I, 1(2), 1–27.

[54] Cheng, F., Chen, X., Xiang, C., Liu, Z., & Wang, Y. (2020). Fluorescence-based high-throughput screening system for R - ω -transaminase engineering and its substrate scope extension.

[55] Mussardo, G. (2019). No Title No Title. Statistical Field Theor, 53(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

[56] Wang, T., Zhao, B., & Foehr, E. D. (2019). Engineering a Long Lasting Tethered, Multimeric Human Growth Hormone Protein to Improve Pharmacokinetic Half-Life and Potency. Journal of Proteomics & Bioinformatics, 12(3), 56–60. https://doi.org/10.35248/0974-276x.19.12.497