International Journal of Peptide Research and Therapeutics (2021) 27:1315–1329
https://doi.org/10.1007/s10989-021-10170-7

iDRP-PseAAC: Identification of DNA Replication Proteins Using General PseAAC and Position Dependent Features

Arqam Amin[1] · Muhammad Awais[1] · Shalini Sahai[2] · Waqar Hussain[3,4] · Nouman Rasool[4]

Accepted: 18 January 2021 / Published online: 8 February 2021

Abstract

DNA replication is one of the specific processes to be considered in all the living organisms, specifically eukaryotes. The prevalence of DNA replication is significant for an evolutionary transition at the beginning of life. DNA replication proteins are those proteins which support the process of replication and are also reported to be important in drug design and discovery. This information depicts that DNA replication proteins have a very important role in human bodies, however, to study their mechanism, their identification is necessary. Thus, it is a very important task but, in any case, an experimental identification is time-consuming, highly-costly and laborious. To cope with this issue, a computational methodology is required for prediction of these proteins, however, no prior method exists. This study comprehends the construction of novel prediction model to serve the proposed purpose. The prediction model is developed based on the artificial neural network by integrating the position relative features and sequence statistical moments in PseAAC for training neural networks. Highest overall accuracy has been achieved through tenfold cross-validation and Jackknife testing that was computed to be 96.22% and 98.56%, respectively. Our astonishing experimental results demonstrated that the proposed predictor surpass the existing models that can be served as a time and cost-effective stratagem for designing novel drugs to strike the contemporary bacterial infection.

Keywords DNA replication · Replication proteins · PseAAC · 5-Steps rule · Prediction

Introduction

DNA replication is one of the specific processes in eukaryotes. The prevalence of DNA replication should be a significant evolutionary transition at the beginning of life. By replicating DNA content, organisms can pass genetic information on to future generations (Fragkos et al. 2015; Kurat et al. 2017; Vaz et al. 2016). Mutations during the reproduction process allow the population to evolve and adapt. The central importance of DNA replication for such important processes in life makes the development of the DNA replication mechanism more important for understanding the evolution of life (Wang et al. 2004).

DNA replication is a biological process that produces two identical copies of DNA from the original DNA molecule. DNA replication occurs in all living organisms. The cells have the characteristic that requires DNA replication to be carried out (Beattie et al. 2017; Yeeles et al. 2017). Double helix DNA consists of two integrated branches. These strings are separated during the copy process. Next, each strand of the original DNA molecule functions as a template and generates its counterpart. This is a process known as semi-conservative iteration. Because of the semi-conservative replication, the new coil is composed of both the original DNA strand and the newly synthesized strand. Cell error correction and error-checking mechanisms ensure almost complete commitment to DNA replication (Fragkos et al. 2015; Kurat et al. 2017).

In cells, DNA replication begins at a specific site or origin of replication in the genome. The formation of DNA and the synthesis of new strands, called Helis enzyme uptake, results in the appearance of repetitive spinous processes that grow in both directions from the original direction. Many proteins

✉ Muhammad Awais
ta.awaisajaz@gmail.com

[1] Department of Information System, University of Management and Technology, Lahore, Pakistan

[2] Cloud Fusion Lab, Chicago, IL, USA

[3] National Center of Artificial Intelligence, Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan

[4] Center for Professional & Applied Studies, Lahore, Pakistan

bind to the replication fork and initiate and maintain DNA synthesis. Importantly, DNA polymerase synthesizes new strands by adding complementary nucleotides to each strand. DNA replication occurs in the synthesis (Aze et al. 2016).

In the area of vaccine development, DNA replication proteins are of great interest being immunologically active for inducing the immune system (Hamzeh-Mivehroud et al. 2013). DNA replication proteins are reported to be very potentially active targets against antimicrobial agents (Eijk et al. 2017). This information depicts that DNA replication proteins have a very important role in human bodies. Transcription termination factor is one of the important examples of DNA replication proteins (Fig. 1). However, to study their mechanism, their identification is necessary. Thus, it is a very important task but, in any case, an experimental identification is time-consuming, highly-costly and laborious. To cope with this issue, a computational methodology is required for prediction of these proteins (Jiang et al. 2016; Li et al. 2016), however, no prior method exists.

Computational methodologies or bioinformatic tools and approaches have extensively been used to provide insight and valuable information at the molecular as well as protein level. These approaches include structural bioinformatics tools (Chou 2004), molecular packing (Chou et al. 1988), molecular docking (Gao et al. 2007; Wang et al. 2000; Zheng et al. 2007; Li et al. 2007; Zhang et al. 2006), protein subcellular location prediction (Chou and Shen 2007a,b), membrane proteins prediction with their types (Chou and Shen 2007b), prediction of classes and subclasses of functional enzymes (Chou and Shen 2007b), approaches for protease cleavage sites and single peptides prediction (Chou and Shen 2007a,b; Shen and Chou 2008) and QSAR models to predict specific activities of peptides and proteins for drug designing. The bioinformatic analysis provides by these computational approaches has developed remarkable advances for better understanding of the nucleic acids and proteins, their interactions and mode of actions that helped in the development of a wide variety of novel drugs to target extensive range of microbial infections.

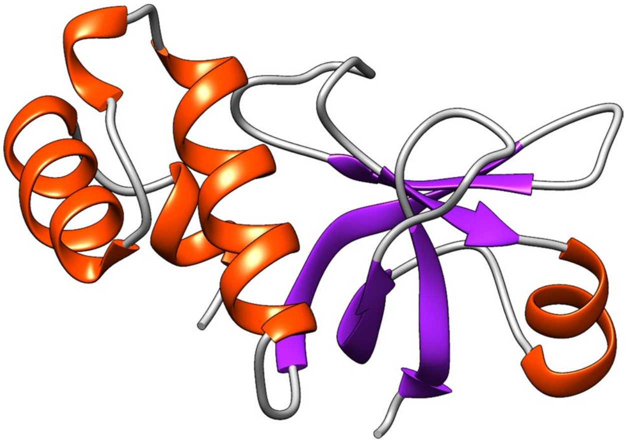

Fig. 1 Transcription termination factor (PDB ID: 2A8V). Red denotes α-helices, purple denotes β-sheets while white denotes random coil in tertiary structure (Color figure online)

Bioinformatic analysis of proteomics revealed the fundamental requirements of the discrimination between DNA replication protein and non-DNA replication proteins. Numerous algorithms have been proposed over the past decades for the predicting structure of proteins (Xiao et al. 2008) protein classification (Li and Li 2008), proteins superfamily, family and subfamily classes (Li and Li 2008; Cai et al. 2005; Zhou et al. 2007), prediction of protein subnuclear and subcellular localization (Ding and Zhang 2008; Jiang et al. 2008; Lin 2008; Lin et al. 2008) and other protein cellular attributes (Chou 2001a). These protein attributes predicting algorithms include support vector machine; SVM) (Lin 2008; Lin et al. 2008; Chen et al. 2008), K-nearest neighbor; KNN) (Shen and Chou 2005a,b; Yan et al. 2008) and Fisher discriminant classifier (Ding et al. 2009).

The pseudo amino acid composition (PseAAC) has been demonstrated to efficiently improve the calibre of protein prediction by presenting a distinct model of protein-peptide sequence deprived of lacking the information of sequence order of protein (Li and Li (2008); Chou 2001a; Du et al. 2014). For statistical prediction, the efficiency of a predictor is most often examined by the number of cross-validation methods including independent dataset test, self-consistency testing, subsampling test, K-fold cross-validation test and jackknife test (Shen and Chou 2008; Liu et al. 2016a,b; Butt et al. 2016, 2017). The present study was conducted to construct a novel computational predictor for predicting DNA replication proteins and for discriminating DNA replication proteins with non-DNA replication proteins. This model will provide beneficial and worthy information for successful prediction of DNA replication proteins. The Chou's PseAAC can integrate the chief attributes of the composition of an amino acid as well as the correlation of sequence order. This sequence-based statistical predictor operates based on the following five prime rules i.e. 5-step rule (Chou 2020a,b,c,d,e,f; Fang et al. 2020; Lin et al. 2020; Liu and Chou 2020; Lu and Chou 2020; Shao and Chou 2020; Shao et al. 2020; Xu et al. 2020; Zhang et al. 2020) which is (i) Construction or selection of an effective benchmark dataset for training and testing the sequence-based statistical predictor, (ii) Formulation of the effectual biological sequence tasters with an operative measured expression to accurately replicate the intrinsic relation of the biological sequence with the target to be prophesied, (iii) Development of a productive and efficacious algorithm for operating the prediction, (iv) Execution of persuasive cross-validation trials to factually assess the projected precision of the predictor, and (v) The inception of a comprehensible and foolproof web-server regarding the predictor and to ensure its receptiveness and accessibility to the public.

Materials and Methods

In this section, the first three steps of the 5-steps prime rule are being addressed. Proposed Methodology is being illustrated in the flowchart in Fig. 2.

Organizing the particular and precise set of data for testing and training is the utmost priority for creating a potent predictor. Flawed, imprecise or fallacious benchmark dataset consequences in deceitful and unreliable predictor training that leads to false verification and validation of the particular data. That is why the collection of the exhaustive and non-redundant dataset is of prime importance. This study implies the construction of computational model by retrieving the protein sequences and formation of comprehensive benchmark dataset from a renowned online protein database such as UniProt during the first phase followed by the extraction of feature vector comprising of pertinent features in numerals during the second phase. During the next phase, the retrieved features were then trained with the help of a neural network for attaining convergence. For this proposed trained predictor, feature vectors serve as an input to predict output that specifies whether the particular sequence is of DNA replication protein or non-DNA replication protein. In the end, various verification tests were employed for the predictor model followed by its validation on several test datasets to establish the accuracy and preciseness of the predictor model so that the possibility and precision of the trained model could be illustrated in Fig. 3.

Data Collection

There were 5301 DNA replication proteins and 5250 non-DNA replication proteins retrieved from UniProt databases. “DNA replication protein” as a keyword was used for retrieval of the protein sequence. Scrupulous data collection was ensured by excluding the ambiguous sequences and only the sequences containing specific attributes were retrieved. These specific attributes include annotation of the particular sequence with the terms viz “DNA replication” by similarity, probable, potential, fragment. Moreover, these were all reviewed sequences and non-reviewed sequences were not considered. We searched all DNA replication proteins from UniProt, irrespective of length, to make dataset rich and discriminant, so that the model trained could be more dynamic. Similarly, for non-DNA replication proteins as well, reviewed sequences were retrieved. The present study dealt with the removal of redundant proteins by using CD-HIT process and proteins sequence identity as the cutoff was 60%. 60% cutoff depicts that all sequences, which showed similarity more than 60% were excluded from the dataset to reduce redundancy in dataset and overfitting of the model. The reason for choosing this threshold was that it is supported by various previously reported studies (Shao and Chou 2020; Shao et al. 2020; Khan et al. 2019a,b; Jia et al. 2019; Ilyas et al. 2019; Hussain et al. 2019a,b; Feng et al. 2019; Cui et al. 2019; Awais et al. 2019). Clusters were formed comprising of 5101 non-redundant DNA replication proteins and 5227 non-redundant non-DNA replication proteins.

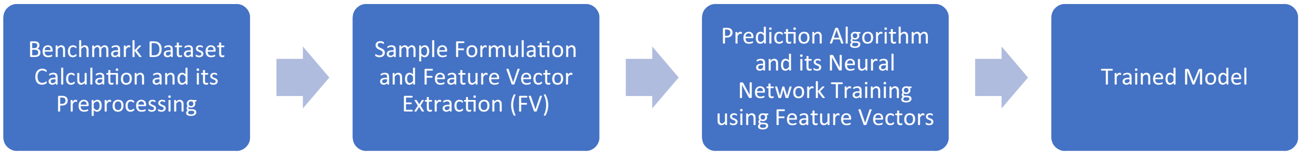
Benchmark Dataset Calculation and its Preprocessing
Sample Formulation and Feature Vector Extraction (FV)
Prediction Algorithm and its Neural Network Training using Feature Vectors
Trained Model

Fig. 2 Linear flow diagram of proposed methodology

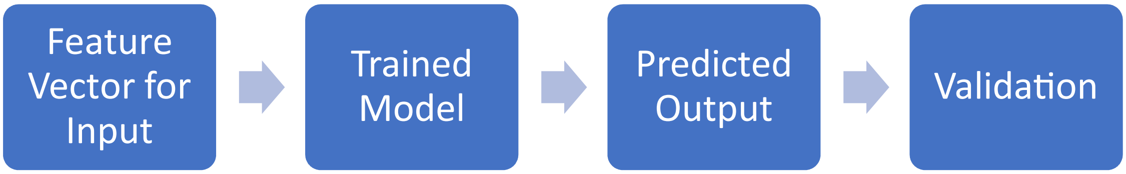

Fig. 3 Linear flow diagram of trained model

Feature Vector Construction

The precise order of incorporation of the amino acid sequence into the polypeptide chain of the protein defines the characteristic properties of that specific protein encoded by the specific gene. Attributes of a particular protein can be altered in terms of structure or function as a consequence of amino acid mutations or due to the presence or absence of a signal amino acid in the sequence of the particular gene. The relative placing of integral amino acid residues is far more substantial than the amino acid composition that significantly affects the behaviour of the protein. The reason behind the eminent alteration of the characteristic biophysical properties of the protein lies at the small variation in the relative positioning of amino acids (Liu et al. 2016a). These facts reinforce the development of mathematical and computational models to retrieve information of the characteristic features from the protein's primary sequence with the regard of the relative positioning of the amino acids rather than the constituents of the proteins.

Statistical Moments of the Primary Structure (SM)

Statistical moments referred to as a measure of data collection quantitatively. Several orders of moments depict numerous attributes of the data. The moments delineate the evaluation of the data size as well as its orientation and eccentricity. Varied moments have been constructed on the basis of renowned distribution and polynomials functions. There are various eminent moments to explicate the anticipated problem viz. raw, central and Hahn moments. Raw moments reckon variance, mean and asymmetry of the distribution of probability irrespective of scale invariance or location invariance (Butt et al. 2016; Khan et al. 2014). Central moments along with the information similar to the raw moments; computed along the centroid of the data with respect to scale invariance or location invariance (Butt et al. 2016; Khan et al. 2014). Hahn moments are computed on the basis of Hahn polynomials irrespective of scale invariance or location invariance. The attribute of moment selection depends on the susceptive of the moment towards sequence ordered information which is of prime importance. Therefore, scale-invariant moments are ignorable so these moments were evaded. The computed values from all these methods elucidate data differently. Moreover, the discrepancy between the computed values for the moments of random dataset implicit discrepancies in the features of the data source (Butt et al. 2017).In the present study, the bi-dimensional version of these moments was implied by transforming the single-dimensional primary sub-sequence into bi-dimensional notation.

In supposition, the sequence of the protein or sub-sequence 'R' can be denoted by:

$$R = (\alpha 1, \alpha 2, \alpha 3, \ldots, \mathrm{AK}) \tag{1}$$

where α_i represents the ith residue of an amino acid in a primary sub-sequence comprising of k residues, again suppose,

$$n = \left[\sqrt{\mathrm{K}}\right] \tag{2}$$

For accommodating all the amino acid residues of the protein R, A matrix R' is composed of dimension m*m

$$\mathrm{R}' = \left\{ \begin{matrix} \gamma_{11} & \gamma_{2} & \cdots & \gamma_{1\mathrm{n}} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2\mathrm{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{\mathrm{n}1} & \gamma_{n2} & \cdots & \gamma_{\mathrm{nn}} \end{matrix} \right\} \tag{3}$$

As the bi-dimensional matrix R' belongs to the primary structure R, therefore, the matrix R can be transformed into R' using a mapping function ω as follow:

$$\omega\left(\alpha_m\right) = \beta_{ij} \tag{4}$$

whereas, $i = \mathrm{m/n} + 1$ & $j = m \bmod n$.

The matrix of 2D R' contents is used to compute the moments to the degree 3. Therefore, the raw moments can be calculated by the equation as follow:

$$Mij = \sum\nolimits_{R=1}^{m} \sum\nolimits_{q=1}^{m} \mathrm{R}^{i}\, \mathrm{q}^{\,j} \beta_{Rq} \tag{5}$$

where i + j represents the moments, order computed till order 3 and can be expressed as $ƙ_{00}$, $ƙ_{01}$, $ƙ_{10}$, $ƙ_{11}$, $ƙ_{12}$, $ƙ_{21}$, $ƙ_{30}$ and $ƙ_{03}$.

The centroid of the data corresponds to the point where the data is consistently dispersed in all directions regarding the weighted average and it can be simply calculated followed by the computation of raw moments. It is expressed as a point $\bar{\upsilon}$, ỹ where

$$\overline{v} = \mathcal{K}_{10}/\mathcal{K}_{00} \text{ and } \tilde{y} = \mathcal{K}_{01}/\mathcal{K}_{00} \tag{6}$$

The central moment was computed using centroid with the help of the following equation:

$$\overline{e}_{ig} = \sum_{R=1}^{m} \sum_{q=1}^{m} \left(R - \overline{v}\right)^{i} (R - \tilde{y})^{j} \beta_{Rq} \tag{7}$$

A square matrix notation R' was formed by transforming one-dimensional notation R. This transformation exhibits greater dividend so as Hahn moments to be calculated on such a level dimensional organization of data as it requires a square matrix as a bi-dimensional input data. Being an orthogonal entity, discrete Hahn moments have a reversible property that renders the reconstruction of the original data with the utilization of inverse functions of Hahn moments. Therefore, it is evident that the computed moments conserved the positional and conformational information of a primary sequence.

Let the Hahn polynomial order of' to be specified by the following equation:

$$h_m^{u,v}(d,Y) = (Y + \mathcal{F} - 1)_m(Y - 1)_m \times \sum_{k=0}^{m}(-1)\kappa^k \frac{(-m)_\kappa(-d)_\kappa(2Y + u + v + -1)_\kappa}{(Y + v - 1)_\kappa(Y - 1)_\kappa}\frac{1}{\kappa!} \tag{8}$$

The pochhammer symbols used in the above equation was generalized as:

$$(\mathcal{F})_k = \mathcal{F} \cdot (\mathcal{F} + 1) \ldots (\mathcal{F} + \kappa - 1) \tag{9}$$

Gamma operator was used to abridging the expression as follow:

$$(\mathcal{F})_k = \frac{\Gamma(\mathcal{F} + k)}{\Gamma(\mathcal{F})} \tag{10}$$

With the help of weighting function and square norm, the raw values of Hahn moments were ascended as below:

$$\tilde{h}_m^{u,v}(d,Y) = h_m^{u,v}(d,Y), \sqrt{\frac{p(r)}{s_m^2}}, m = 0,1 \ldots \mathrm{Y} - 1 \tag{11}$$

whereas;

$$P(d) = \frac{(u + d + v)(v + d + 1)(u + v + d + 1)_Y}{(u - v - 2\mathrm{r} - 1)m!(Y - d - 1)!} \tag{12}$$

Finally, for bi-dimensional discrete data matrix, the orthogonal normalized Hahn moments were computed by the following equation:

$$L_{ig} = \sum\nolimits_{q=0}^{Y-1}\sum\nolimits_{p=0}^{Y-1}\beta_{ig}\tilde{h}_m^{u,v}(q,Y)\tilde{h}_m^{u,v}(R,Y), n, m = 0, 1, \ldots Y - 1 \tag{13}$$

For every primary sequence, bi-dimensional raw, central and Hahn moments were computed up to third order and afterwards, these were combined with the miscellany feature vector.

Position Relative Incidence Matrix (PRIM)

The basic information of sequence order encoded into the primary sequence of the protein and relative positioning information of amino acid residues, being a chief paradigm exhibits the foundation of any mathematical model for the prediction of attributes and characteristics of the protein. The quantization of the relative position of an amino acid within the polypeptide chain is also of prime importance. Position relative incidence matrix (PRIM), generated by the elements with 20×20 dimensions, signifies the comprehensive information concerning relative positioning of amino acid residues within the protein's polypeptide chain.

$$S_{PRIM} = \left\{\begin{matrix} M_{1\to1} & M_{1\to2\cdots} & M_{1\to j\cdots} & M_{1\to20} \\ M_{2\to1} & M_{2\to2\cdots} & M_{2\to j\cdots} & M_{2\to20} \\ M^{\vdots}_{i\to1} & M^{\vdots}_{i\to2\cdots} & M^{\vdots}_{i\to j\cdots} & M^{\vdots}_{i\to20} \\ M^{\vdots}_{N\to1} & M^{\vdots}_{N\to2\cdots} & M^{\vdots}_{N\to J\cdots} & M^{\vdots}_{N\to20} \end{matrix}\right\} \tag{14}$$

An element of $_{i\to j}$ represents the summation of the relative position of *j*th residue concerning the first incidence of the *i*th residue. PRIM yielded a huge number of coefficients which were then further reduced up to 24 elements through computing the statistical moments employing PRIM as the input.

Reverse Position Relative Accumulative Matrix (RPRIM)

The fact related to proficiency and the precision of the machine learning algorithm lies behind the punctiliousness and the fastidiousness for extracting the utmost pertinent data set. Ambiguous patterns entrenched within data can be understood and uncover with the self-adapting capability of a machine learning algorithm. As PRIM matrix uncovers the information concerning with the relative positioning of amino acid residues within the polypeptide chain of a protein, reverse position relative incidence matrix (RPRIM) was employed to uncover the obscure hidden attributes of the primary sequence of the protein that ultimately extenuate opacities among proteins with apparently identical sequences. Likewise, as PRIM, RPRIM also generated as a 20×20 dimension of elements and yielded 400 coefficients nevertheless in the reverse primary sequence which was further reduced to 24 coefficients by computing the moments. Reverse position relative accumulative matrix is given as follow:

$$S_{PRIM} = \left\{\begin{matrix} Y_{1\to1} & Y_{1\to2\cdots} & Y_{1\to j\cdots} & Y_{1\to20} \\ Y_{2\to1} & Y_{2\to2\cdots} & Y_{2\to j\cdots} & Y_{2\to20} \\ Y^{\vdots}_{i\to1} & Y^{\vdots}_{i\to2\cdots} & Y^{\vdots}_{i\to j\cdots} & Y^{\vdots}_{i\to20} \\ Y^{\vdots}_{N\to1} & Y^{\vdots}_{N\to2\cdots} & Y^{\vdots}_{N\to J\cdots} & Y^{\vdots}_{N\to20} \end{matrix}\right\} \tag{15}$$

Determination of the Frequency Matrix (FM)

The number of times each amino acid residue occurs within the polypeptide chain of the primary sequence of a protein designated as a frequency and frequency matrix was designed to measure the distribution of frequency of an amino acid residue in the sequence. The frequency matrix is given as follow:

$$\xi = \left\{\tau_1,\ \tau_2,\cdots\cdots,\ \tau_{20}\right\} \tag{16}$$

whereas τ_i denotes the frequency of occurrence of ith amino acid residue. The frequency matrix comprehends the information concerning the configuration and conformation of the protein. Moreover, computing the frequency matrix aims at extracting the structural evidence of the protein sequence.

Generation of Accumulative Absolute Position Incidence Vector (AAPIV)

As frequency matrix represents how much specific amino acid residue is frequent, it gives the structural and conformational information but not the relative position of amino acid residues in a protein. Hence, accumulative absolute position incidence vector was computed for the relative position of amino acid and for extracting the composition of the protein. AAPIV constitutes of 20 elements and each element denotes the summation of all the ordinal values for individual amino acid positioned at their corresponding site within the primary sequence of a protein.

AAPIV vector has represented the incidence of particular amino acid residue in the primary sequence and is given by:

$$\alpha^i_{p_1} \ldots \alpha^i_{p_2} \ldots \alpha^i_{p_3} \ldots \alpha^i_{pn} \tag{17}$$

α^i_{pn} signifies the occurrence of specific amino acid residue α^i at positions of $p_1, p_2, p_3 \ldots p_n$. Accordingly, Accumulative absolute position incidence vector is designated as follow:

$$K = \left\{\mu_1, \mu_2, \mu_3 \ldots, \mu_{20}\right\} \tag{18}$$

Henceforth, for an arbitrary ith element, AAPIV can be computed as follow:

$$\mu_i = \sum_{k=1}^{n} P_k \tag{19}$$

Generation of Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)

Reverse accumulative absolute position incidence vector was generated to extract the information of deep and ambiguous pattern about the relative locations of amino acid residues in the protein sequence. RAAPIV is also a 20-element vector and was generated by reversing the primary sequence followed by the extraction of reverse accumulative absolute position incidence vector from the particular reversed sequence and is computed as follow:

$$A = \left\{\eta_1, \eta_2, \eta_3 \ldots, \eta_{20}\right\} \tag{20}$$

Suppose the incidences of a particular amino acid residue in the reversed sequence be represented as follow:

$$\alpha^i_{l_1} \ldots \alpha^i_{l_2} \ldots \alpha^i_{l_3} \ldots \alpha^i_{\mathrm{ln}} \tag{21}$$

where l_1, l_2, l_3, …, l_n signifies the ordinal positions for the occurrence of amino acid residue (α^i) that particular reverse sequence. Therefore, for an arbitrary ith element, reverse accumulative absolute position incidence vector (RAAPIV) can be computed as follow:

$$\eta_i = \sum\nolimits_{k=1}^{n} l_k \tag{22}$$

Neural Network Training

Decision complications can be resolved with an utmost commanding technique referred as a neural network that resembles the human nervous system as the brain grasp and absorbs the environmental information and acts according to the scenario by learning from the circumstances. A neural network has been structured on the analogous code. During the training operation, it acquires characterized input, it projected the judgment for each input and based on knowledge obtained from each input. Two approaches are employed for the training of neural network categorized as supervised and unsupervised. Former comprehends both the inputs and the outputs in which network processing of the input outcomes the desired output while later encompasses the ability to sense the input deprived of external assistance (Fig. 4).

After the accomplishment of the network training, the intrinsic capability of the network enables it to organize the respective input within an adequate level of precision. Reduction of an error is the prime objective during the learning procedure of the neural network which regulates its weights throughout each reiteration to diminish the possibility of an error. It eventually aids to translate into upgraded learning and enhanced precision for predicting the pertinent group of random input.

An artificial neural network is an immensely efficacious approach for the creating supervised classifier for the development of prediction algorithm for DNA replication protein and non-DNA replication proteins. The depths and particulars drawn out from raw data into the feature vector possess a dynamic part. A feature vector is proficient for discerning data to obtain diligent outcomes. The construction of feature vector embraces discriminating attributes including FM, SVV, AAPIV and RAAPIV. It also employs raw, Hahn moments and central moments of PRIM, RPRIM along with the bi-dimensional primary structure. A feature vector consists of a large numeral of coefficients that efficiently

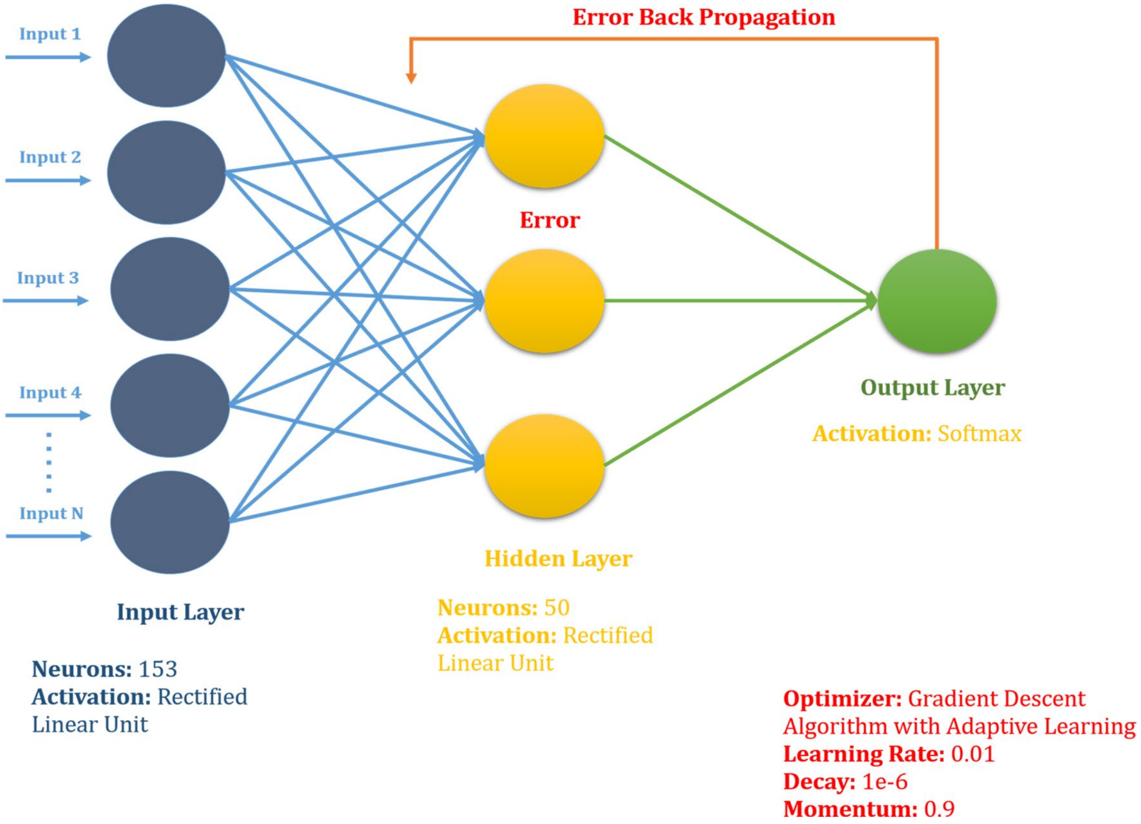

Fig. 4 The architecture of ANN for the proposed prediction model

predict the DNA replication protein and non-DNA replication proteins.

Experimentation and Results

Now, the fourth step of Chou's 5-step rule will be discussed. The estimation of exactness helps in target assessment of another algorithm for prediction. To address this, it is important to think about which sort of measurements are to be utilized and what test technique would be utilized to score those measurements.

Accuracy Estimation Metrics

There are four factors which are used to measure the prediction of DNA replication proteins. The accuracy (Acc) factor used to measure the overall prediction accuracy, the sensitivity (Sn) factor used to check the sensitivity level, Mathew's correlation coefficient (MCC) factor used to measure the overall stability of predictor and the Specificity (Sp) factor used to check the specificity. Lamentably, their regular details as given in ref. (Chen et al. 2007) need instinct and most exploratory researchers feel difficult to comprehend them, especially for the MCC. The traditional definitions of these parameters do not have the instinct and these parameters are viewed as troublesome by the researchers to comprehend (Chou 2001b), in this way, Chou's symbols, changed over by Xu et al. (2013) and Chen et al. (2013), were utilized for estimation of exactness which is given as

$$\begin{cases} Sn = 1 - \frac{\mathbb{N}_{-}^{+}}{\mathbb{N}^{+}} & 0 \le Sn \le 1 \\ Sp = 1 - \frac{\mathbb{N}_{+}^{-}}{\mathbb{N}^{-}} & 0 \le Sp \le 1 \\ Acc = 1 - \frac{\mathbb{N}_{-}^{+}+\mathbb{N}_{+}^{-}}{\mathbb{N}^{+}-\mathbb{N}^{-}} & 0 \le Acc \le 1 \\ MCC = \frac{1-\left(\frac{\mathbb{N}_{-}^{+}}{\mathbb{N}^{+}}+\frac{\mathbb{N}_{+}^{-}}{\mathbb{N}^{-}}\right)}{\left(1+\frac{\mathbb{N}_{+}^{-}-\mathbb{N}_{-}^{+}}{\mathbb{N}^{+}}\right)\left(1+\frac{\mathbb{N}_{-}^{+}-\mathbb{N}_{+}^{-}}{\mathbb{N}^{-}}\right)} & -1 \le MCC \le 1 \end{cases}$$

where $\mathbb{N}^{+}$ shows to the number of real predicted DNA replication proteins, and $\mathbb{N}_{-}^{+}$ shows to the number of real DNA replication proteins which are predicted as non-DNA replication proteins by applying the above matrix algorithm. Essentially, $\mathbb{N}^{-}$ shows to the number of real predicted non-DNA replication proteins, and $\mathbb{N}_{+}^{-}$ shows to the number of real non-DNA replication proteins which are predicted as DNA replication proteins by applying the above matrix algorithm. As indicated by the matrix, *Sn* ends up 1 when$\mathbb{N}_{-}^{+} = 0$. Also, *Sp* winds up 1 when$\mathbb{N}_{+}^{-} = 0$. Most extreme accuracy and *MCC* that is *Acc* = 1 and *MCC* = 1 are accomplished when $\mathbb{N}_{-}^{+} = \mathbb{N}_{+}^{-} = 0$, which represents that no mistaken forecasts have been made concerning the DNA replication protein proteins and non-DNA replication proteins. On the off chance that one has $\mathbb{N}_{-}^{+} = \mathbb{N}_{+}^{-}$, it implies that none of a solitary DNA replication protein in the positive dataset and non-DNA replication protein in negative dataset dishonestly anticipated, and it gives us the MCC = 1 and Acc = 1; if one has $\mathbb{N}_{-}^{+} = \mathbb{N}^{+}$ and $\mathbb{N}_{+}^{-} = \mathbb{N}^{-}$, it implies all the DNA replication proteins in positive the dataset and non-DNA replication proteins in the negative dataset are erroneously anticipated, so it gives us the MCC = − 1 and Acc = 0. Then again, on the off chance that one has $\mathbb{N}_{+}^{-} = \mathbb{N}^{-}/2$ and $\mathbb{N}_{-}^{+} = \mathbb{N}^{+}/2$ then it will give us the MCC = 0 and Acc = 0.5, only an estimate whether it is DNA replication protein or non-repair site. In this way, Eq. (19) gives the clarification of explicitness, affectability, in general exactness, and solidness all the straighter forward and instinctive, especially when one talks about MCC. Subsequently, these matrix equations upgrade the instinct and comprehension of exactness measurements, and numerous specialists are likewise as per this (Feng et al. 2019; Song et al. 2018a,b). The arrangement of above equations can be substantial just for the single-class label framework, for example, genuine/false and for frameworks having progressively various labels, an alternate arrangement of parameters can be utilized which is characterized in (Chou 2013).

Self-consistency Test

Self- consistency test referred to as the ultimate test for the validation of efficiency and efficacy of the prediction model using the test cases by training the data set. The obtained results from the self-consistency test were particularized with the assistance of a confusion matrix. Confusion matrix represents an eminent practice to illustrate the precision of a predictor by pronouncing the prediction results counter to the actual data. True positive (TP) represents the entry denoted as a positive for DNA replication protein whereas False positive (FP) identifies the non-DNA replication protein that has been pointed as a DNA replication protein erroneously by the predictor. True negative (TN) recognizes the non-DNA replication protein while False negative (FN) denotes the DNA replication protein inaccurately marked as a non-DNA replication protein by the model. Values of metrics were estimated by putting the values of accuracy parameters into the above equations. A representation of these proposed parameters by conducting the self-consistency testing results for the *iDRP-PseAAC* is shown in Table 1, while ROC and Confusion Matrix is given in Figs. 5 and 6.

Table 1 Self-consistency testing results for iDRP-PseAAC

Predictor	Accuracy metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iDRP-PseAAC	98.56	97.88	99.27	0.97

Prediction Model Validation

There are numerous approaches used for the validation of a predictor elaborated in the literature by various researches (Butt et al. 2016, 2017). K-fold cross-validation and Jackknife testing represent the utmost reliable method as prediction model validation method. Dataset of characteristic validation process typically comprised of training and test data. Firstly, training data was utilized to train the prediction model. After the model being copiously trained and achieved the convergence, untrained data was used to validate the accuracy of the trained predictor. The validation process is demonstrated in Fig. 7 below. Distinct approaches have been employed by the researches for the validation methods using training or test data. For determining the foolproof performance of a predictor, Jackknife testing and Cross-validation are meticulous and diligent approaches (Liu et al. 2015, 2016a,b).

k-Fold Cross-Validation

Cross-validation is a method to thrive an expectancy for the proposed model as an exemplary method in the absence of validation set. Available data was fragmented into K-folds, which is a constant. In this case, the partitions were disjointed and the process was tested for these partitions although being trained for the rest of the data. The test was iterated K times for each partition. The cross-validation result was reported in terms of the overall average of accuracy in each iteration.

Supposedly, let X be the total number of samples comprising of DNA replication protein and non-DNA replication protein samples given as:

$$X = \{X_1, X_2, X_3 \ldots . X_n\} \tag{23}$$

The dataset was split into the subsets $\mathrm{X_i}$ which are of comparable size of k where $\mathrm{X_i}$ represents an arbitrary DNA replication protein or non-DNA replication protein sample.

$$\cup_{i=1}^{k} \mathrm{X_i} = \mathrm{X} \tag{24}$$

$$\cap_{i=1}^{k} \mathrm{X_i} = \mathrm{X} \tag{25}$$

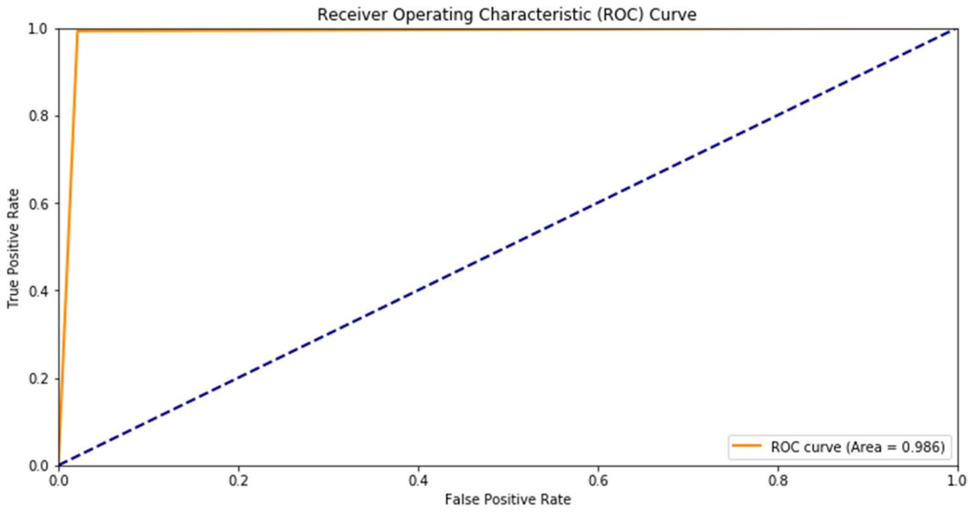

Fig. 5 ROC for iDRP-PseAAC self consistency

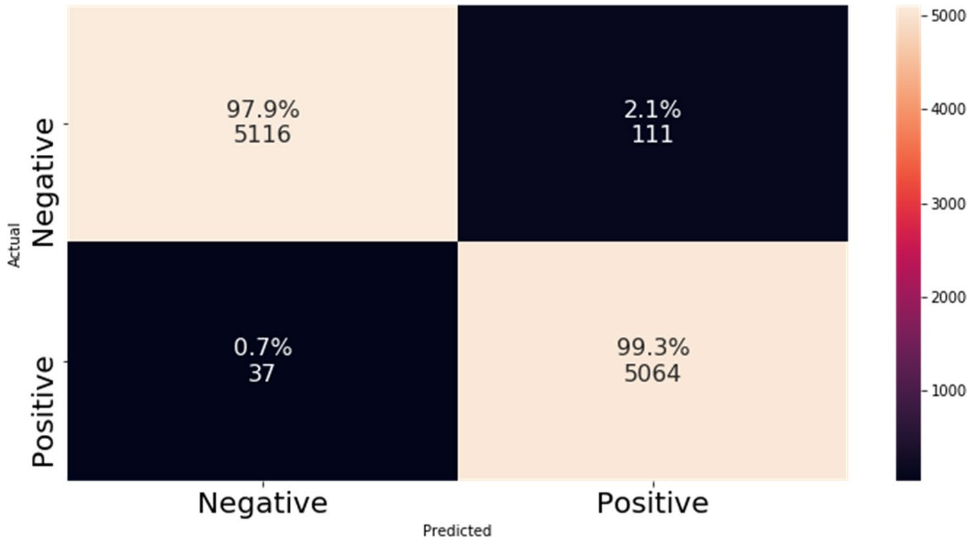

Fig. 6 Conf. matrix for iDRP-PseAAC self consistency

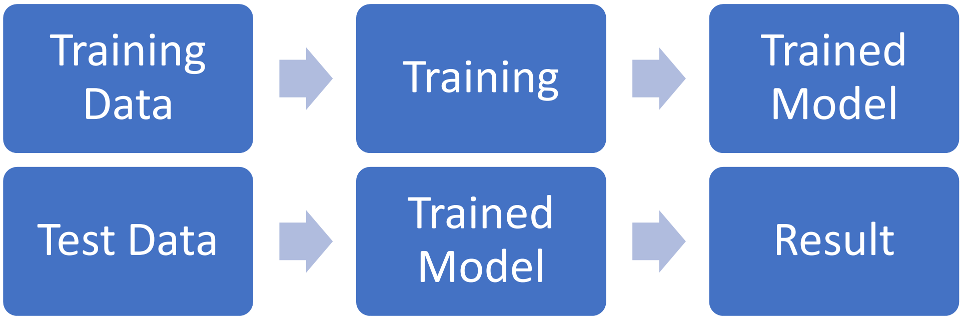

Fig. 7 The validation of prediction model. The validation process is illustrated. The accuracy of the trained model is verified by testing it over partitioned test data

To ensure the comparable sizes of the subsets, these were selected arbitrarily i.e.

$$\left|x_i\right| \cong \left|x_j\right| \tag{26}$$

where X_i and X_j represent discrete arbitrary sets. Elements of X_i were left out during the single iteration and then the predictor was trained on rest of the data. The left-out data was tested using the trained model to compute an accuracy rate R. The mean value for outcomes of k iterations was calculated to compute the overall cross-validation result, R_o.

Table 2 K-fold cross validation testing results for iDRP-PseAAC (average of 10 folds)

Predictor	Accuracy metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iDRP-PseAAC	96.22	95.31	97.15	0.92
Yang et al. (2015)	70.4	68.2	72.6	0.41

$$R_0 = \frac{\sum_{i=1}^{k} R}{k} \tag{27}$$

In the present study, the tenfold cross-validation test was performed separately for DNA replication protein and non-DNA replication protein. Primarily, the test was performed for DNA replication protein and data was divided into test and training data. For the dataset partitioned into 10-folds, a partition was left-out as a test data in each iteration. Afterwards, the neural network was sufficiently trained on the remaining data followed by the simulation to determine its accuracy on test data. Cross-validation test was performed repeatedly on ten datasets encompassing DNA replication protein and non-DNA replication proteins. The average value pronounced the ultimate accuracy of the predictor. The overall accuracy was estimated at 96.22% as demonstrated in Table 2, while the ROC and Confusion Matrix is shown in Figs. 8 and 9. Another study, proposed by Yang et al. (Yang et al. 2015), has been reported previously, which targeted the sequence-based identification of DNA replication proteins, and the

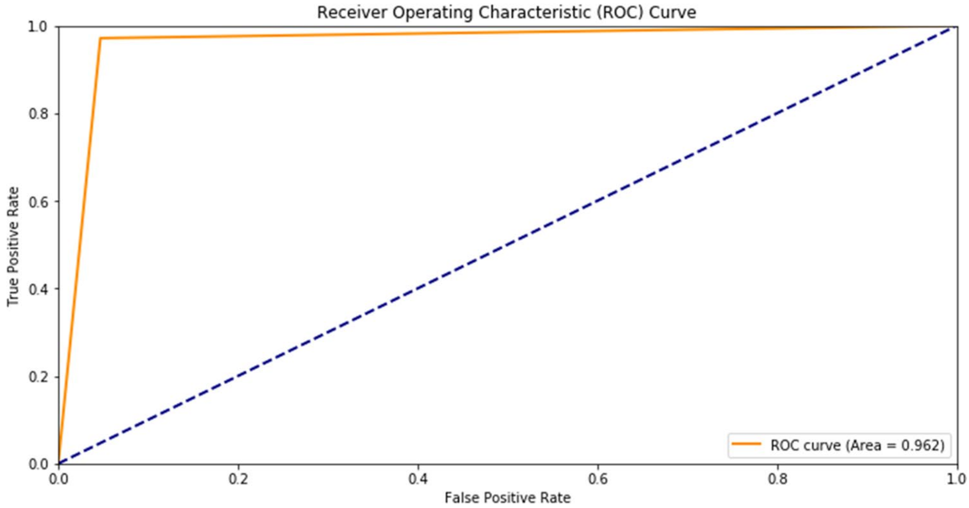

Fig. 8 ROC for iDRP-PseAAC tenfold CV

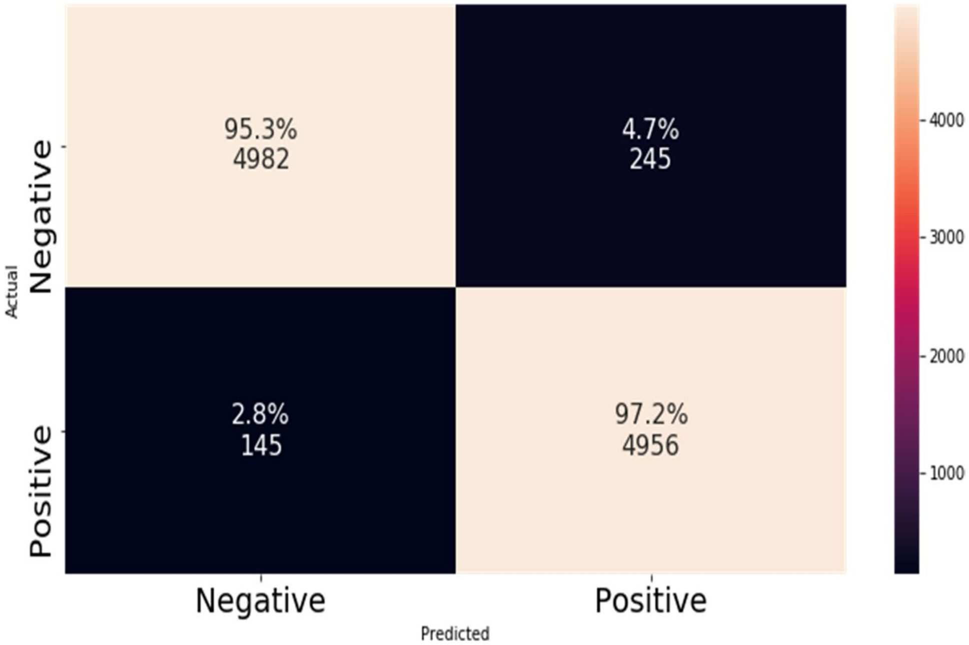

Fig. 9 Conf. Matrix for iDRP-PseAAC tenfold CV

proposed model was validated through tenfold cross-validation. Therefore, the comparison with that study is also presented in Table 2.

Jackknife Testing

Jackknife testing is amongst the most frequently used validation technique. Arbitrarily selected or partitioned datasets are the basis of different validation tests. However, partitioning of the data governs no specific rule. There are several ways to partition the data in a way that certain partitions produce better results whereas certain partition does not give good results. These methods probably fail to produce unique results. Jackknife testing represents the proficient technique capable of producing unique results. It is an iterative technique that computes the accuracy of the model for all variations of the sample of size n − 1. The jackknifing technique train the predictor on left-out data and estimates overall accuracy by meticulously leaving out every observation from a dataset. Eventually, the outcome results of this validation are averaged and produce a unique result for the respective dataset which ultimately alleviates the drawbacks generated by subsampling and data independence.

Supposedly, let S denotes the total sample size comprising of n elements given as follow:

$$S = \{S_1, S_2, S_3 \ldots . S_n\} \tag{28}$$

Let R_i denotes the rate of accuracy for ith iteration of the jackknife test and to compute its value, the dataset leaves out the ith element within the dataset S_i which is given as:

$$S = \{S_1, S_2, S_{3\ldots}., x_{i-1,}x_{i+1,}S_n\}10 \tag{29}$$

The trained neural network along with the feature vector was simulated for all the samples in S_i to compute the accuracy rate, the number of true positives and negatives, as well as false-positive and negatives, were determined. The average of all the result values of R_i was computed as R′.

$$R' = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{30}$$

R' signifies the overall accuracy average of the prediction model and n symbolizes the number of the total observations. The outcome of the model was pragmatically computed for the left-out sample as during each iteration, a sample is excluded out of the training dataset. Iteration was done for entire particular dataset and results were obtained by gathering all predicted results that yielded an accuracy of 98.56%. Jackknife testing results for the *iDRP-PseAAC* is shown in Table 3 while ROC and Confusion Matrix is shown in Figs. 10 and 11.

Table 3 Jackknife testing results for iDRP-PseAAC

Predictor	Accuracy metrics			
	Acc (%)	Sp (%)	Sn (%)	MCC
iDRP-PseAAC	98.56	49.54	49.03	0.0

Jackknife testing is an iterative methodology that calculates the precision of the predictor for all variations of the population of S_{N-1}.

Webserver

This area explains the last step of Chou's 5-steps rule (Liu et al. 2016c; Zhang et al. 2016; Chou 2011) which is the improvement of a web server for the simplicity of clients and easy to understand, as shown by the different examiner in some ongoing publications, easy to understand and freely available web-servers speak to the future heading for growing increasingly helpful prediction strategies and computational analyses tools. They have fundamentally improved the effects of computational science on restorative science (Chou 2015), driving medicinal science into an extraordinary upheaval (Cheng et al. 2017). In this manner, one shall make efforts to develop a webserver for the expectation technique detailed in this paper; until further notice, its independent code is accessible at idrp-server.herokuapp.com/which is created utilizing Django 2.0.7. For a neural system, the sci-unit neural system was utilized with Theano 1.0.0 at the backend. Underneath, the well-ordered manual for the utilization of the webserver is discussed.

Phase 1

First, open the URL of web server idrp-server.herokuapp.com/and then see a top header menu which has four tabs i.e. Home, Prediction, About and Supplementary Data. An overview of DNA replication proteins defines in the Home tab. The DNA replication proteins prediction portal shows in Prediction tab. The reference of relevant paper and its information shows in About tab. The beneficial Data provides for download in the Supplementary Data tab. You just click the Prediction tab to predict the proteins.

Phase 2

In prediction tab, one can input the protein sequence in an empty textbox and then click the Submit button for getting the results. The outcomes will show up on the next window after some time, which relies upon the length of the input.

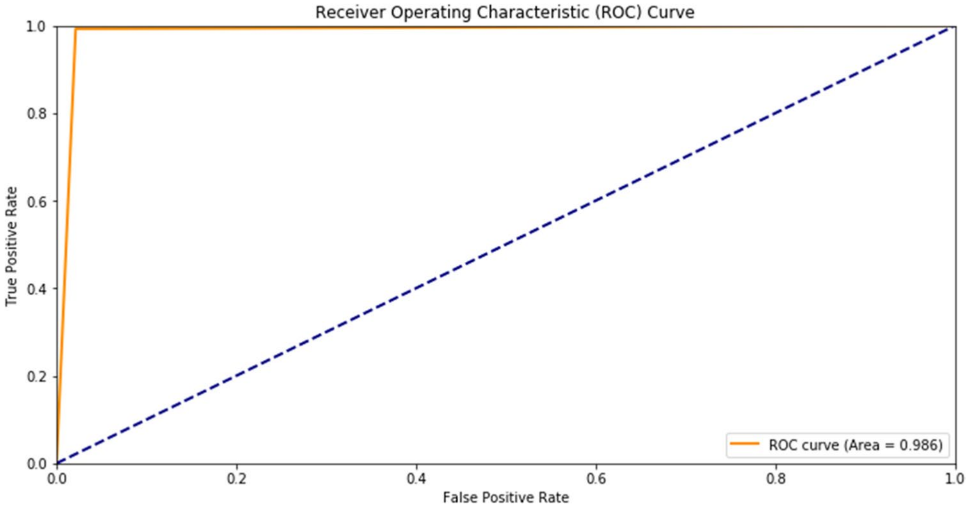

Fig. 10 ROC for iDRP-PseAAC LOO CV

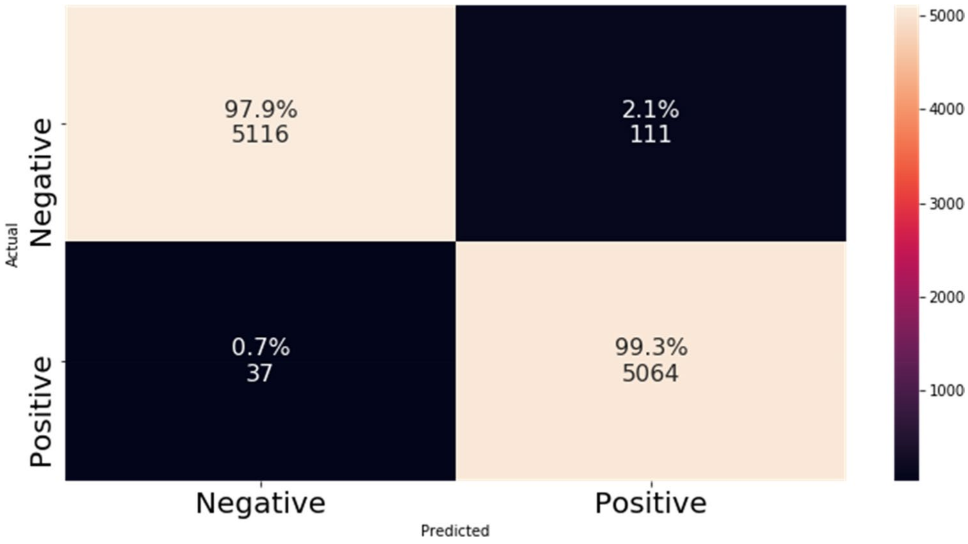

Fig. 11 Confusion Matrix for iDRP-PseAAC LOO CV

Phase 3

To see the relevant paper and algorithm which can be used to develop this server in About tab. One can also see the citation of the relevant paper in this tab.

Phase 4

One can download the relevant or supplementary dataset for future experiments in Supplementary Data tab.

Conclusion

Discrimination of DNA replication protein from the non-DNA replication proteins is a crucial requisite to study the mechanism of these proteins. The present study was conducted to predict the given polypeptide sequences as DNA replication protein or non-DNA replication protein based on fundamental steps of Chou's. Features were calculated by incorporating statistical and position relative features into the amphiphilic pseudo amino acid composition. The results computed from the proposed predictor was validated by employing self-consistency testing, jackknife testing and cross-validation approach. The overall accuracy of the predictor was depicted by using exemplary metrics presenting the high accuracy for the model. Stupendous experimental results demonstrated that the proposed predictor is an accurate and precise approach to conduct further researches as well as it presents time and cost-effective strategy for identifying DNA replication proteins.

Author Contributions All authors contributed equally.

Funding No funding was received for this study.

Data Availability Available at idrp-server.herokuapp.com/.

Code Availability Webserver available at idrp-server.herokuapp.com/.

Compliance with Ethical Standards

Conflict of interest The author declare no conflicts of interest.

References

Awais M, Hussain W, Khan YD, Rasool N, Khan SA, Chou K-C (2019) iPhosH-PseAAC: identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2019.2919025

Aze A, Sannino V, Soffientini P, Bachi A, Costanzo V (2016) Centromeric DNA replication reconstitution reveals DNA loops and ATR checkpoint suppression. Nat Cell Biol 18(6):684

Beattie TR, Kapadia N, Nicolas E, Uphoff S, Wollman AJ, Leake MC, Reyes-Lamothe R (2017) Frequent exchange of the DNA polymerase during bacterial chromosome replication. Elife 6:e21763

Butt AH, Khan SA, Jamil H, Rasool N, Khan YD (2016) A prediction model for membrane proteins using moments based features. BioMed Res Int. https://doi.org/10.1155/2016/8370132

Butt AH, Rasool N, Khan YD (2017) A treatise to computational approaches towards prediction of membrane protein and its subtypes. J Membr Biol 250(1):55–76

Cai Y-D, Zhou G-P, Chou K-C (2005) Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234(1):145–149

Chen J, Liu H, Yang J, Chou K-C (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33(3):423–428

Chen Y-Z, Tang Y-R, Sheng Z-Y, Zhang Z (2008) Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinform 9(1):101

Chen W, Feng P-M, Lin H, Chou K-C (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41(6):e68–e68

Cheng X, Xiao X, Chou K-C (2017) pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol BioSyst 13(9):1722–1727

Chou KC (2001a) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinform 43(3):246–255

Chou K-C (2001b) Using subsite coupling to predict signal peptides. Protein Eng 14(2):75–79

Chou K-C (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11(16):2105–2134

Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273(1):236–247

Chou K-C (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol BioSyst 9(6):1092–1100

Chou K-C (2015) Impacts of bioinformatics to medicinal chemistry. Med Chem 11(3):218–234

Chou K-C (2020a) The most important ethical concerns in science. Nat Sci 12(2):35–36

Chou K-C (2020b) The problem of Elsevier series journals online submission by using artificial intelligence. Nat Sci 12(2):37–38

Chou K-C (2020c) Other mountain stones can attack jade: the 5-steps rule. Nat Sci 12(3):59–64

Chou K-C (2020d) Using similarity software to evaluate scientific paper quality is a big mistake. Nat Sci 12(03):42

Chou K-C (2020e) Proposing 5-steps rule is a notable milestone for studying molecular biology. Nat Sci 12(03):74

Chou K-C (2020f) The development of Gordon life science institute: its driving force and accomplishments. Nat Sci 12(4):202–217

Chou K-C, Shen H-B (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360(2):339–345

Chou K-C, Shen H-B (2007b) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357(3):633–640

Chou K-C, Maggiora GM, Némethy G, Scheraga HA (1988) Energetics of the structure of the four-alpha-helix bundle in proteins. Proc Natl Acad Sci 85(12):4295–4299

Cui X, Yu Z, Yu B, Wang M, Tian B, Ma Q (2019) UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. Chemom Intell Lab Syst 184:28–43

Ding Y-S, Zhang T-L (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recogn Lett 29(13):1887–1892

Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept Lett 16(4):351–355

Du P, Gu S, Jiao Y (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15(3):3495–3506

Fang L, Wang X, Lai Z, Zhang D, Wu M, Pan Z, Wang L, Tang K, Qian D, Huang Z (2020) Reveal the molecular principle of coronavirus disease 2019 (COVID-19). Sci. Program 1(4)

Feng P, Yang H, Ding H, Lin H, Chen W, Chou K-C (2019) iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 111(1):96–102

Fragkos M, Ganier O, Coulombe P, Méchali M (2015) DNA replication origin activation in space and time. Nat Rev Mol Cell Biol 16(6):360

Gao W-N, Wei D-Q, Li Y, Gao H, Xu W-R, Li A-X, Chou K-C (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3(3):221–226

Hamzeh-Mivehroud M, Alizadeh AA, Morris MB, Church WB, Dastmalchi S (2013) Phage display as a technology delivering on the promise of peptide drug discovery. Drug Discov Today 18(23–24):1144–1157

Hussain W, Khan YD, Rasool N, Khan SA, Chou KC (2019a) SPrenylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. J Theor Biol 468:1–11. https://doi.org/10.1016/j.jtbi.2019.02.007

Hussain W, Khan YD, Rasool N, Khan SA, Chou KC (2019b) SPalmitoylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. Anal Biochem 568:14–23. https://doi.org/10.1016/j.ab.2018.12.019

Ilyas S, Hussain W, Ashraf A, Khan YD, Khan SA, Chou K-C (2019) iMethylK-PseAAC: improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou's 5-steps rule. Curr Genom 20(4):275–292

Jia J, Li X, Qiu W, Xiao X, Chou K-C (2019) iPPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC. J Theor Biol 460:195–203

Jiang X, Wei R, Zhao Y, Zhang T (2008) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34(4):669–675

Jiang L, Zhang J, Xuan P, Zou Q (2016) BP neural network could help improve pre-miRNA identification in various species. BioMed Res Int . https://doi.org/10.1155/2016/9565689

Khan YD, Khan NS, Farooq S, Abid A, Khan SA, Ahmad F, Mahmood MK (2014) An efficient algorithm for recognition of human actions. Sci World J. https://doi.org/10.1155/2014/875879

Khan YD, Jamil M, Hussain W, Rasool N, Khan SA, Chou KC (2019a) pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments. J Theor Biol 463:47–55. https://doi.org/10.1016/j.jtbi.2018.12.015

Khan YD, Amin N, Hussain W, Rasool N, Khan SA, Chou K-C (2019b) iProtease-PseAAC (2L): a two-layer predictor for identifying proteases and their types using Chou's 5-step-rule and general PseAAC. Anal Biochem 588:113477

Kurat CF, Yeeles JT, Patel H, Early A, Diffley JF (2017) Chromatin controls DNA replication origin selection, lagging-strand synthesis, and replication fork rates. Mol Cell 65(1):117–130

Li F-M, Li Q-Z (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15(6):612–616

Li Y, Wei D-Q, Gao W-N, Gao H, Liu B-N, Huang C-J, Xu W-R, Liu D-K, Chen H-F, Chou K-C (2007) Computational approach to drug design for oxazolidinones as antibacterial agents. Med Chem 3(6):576–582

Li D, Ju Y, Zou Q (2016) Protein folds prediction with hierarchical structured SVM. Curr Proteom 13(2):79–85

Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252(2):350–356

Lin H, Ding H, Guo F-B, Zhang A-Y, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15(7):739–744

Lin W, Xiao X, Qiu W, Chou K-C (2020) Use Chou's 5-steps rule to predict remote homology proteins by merging grey incidence analysis and domain similarity analysis. Nat Sci 12(03):181

Liu X-X, Chou K-C (2020) pLoc_Deep-mGneg: predict subcellular localization of gram negative bacterial proteins by deep learning. Adv Biosci Biotechnol 11(5):141–152

Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C (2015) Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE 10(3):e0121501

Liu B, Fang L, Liu F, Wang X, Chou K-C (2016a) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn 34(1):223–235

Liu B, Wang S, Dong Q, Li S, Liu X (2016b) Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. IEEE Trans Nanobiosci 15(4):328–334

Liu B, Long R, Chou K-C (2016c) iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics 32(16):2411–2418

Lu Z, Chou K-C (2020) iATC_Deep-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals by deep learning. Adv Biosci Biotechnol 11(5):153–159

Shao Y-T, Chou K-C (2020) pLoc_Deep-mAnimal: a novel deep CNN-BLSTM network to predict subcellular localization of animal proteins. Nat Sci 12(5):281–291

Shao Y-T, Liu X-X, Lu Z, Chou K-C (2020) pLoc_Deep-mPlant: predict subcellular localization of plant proteins by deep learning. Nat Sci 12(5):237–247

Shen H, Chou K-C (2005a) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334(1):288–292

Shen H-B, Chou K-C (2005b) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337(3):752–756

Shen H-B, Chou K-C (2008) HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. Anal Biochem 375(2):388–390

Song J, Wang Y, Li F, Akutsu T, Rawlings ND, Webb GI, Chou K-C (2018a) iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform. https://doi.org/10.1093/bib/bby028

Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou K-C, Webb GI (2018b) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol 443:125–137

van Eijk E, Wittekoek B, Kuijper EJ, Smits WK (2017) DNA replication proteins as potential targets for antimicrobials in drug-resistant bacterial pathogens. J Antimicrob Chemother 72(5):1275–1284. https://doi.org/10.1093/jac/dkw548

Vaz B, Popovic M, Newman JA, Fielden J, Aitkenhead H, Halder S, Singh AN, Vendrell I, Fischer R, Torrecilla I (2016) Metalloprotease SPRTN/DVC1 orchestrates replication-coupled DNA-protein crosslink repair. Mol Cell 64(4):704–719

Wang I-N, Smith DL, Young R (2000) Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol 54(1):799–825

Wang X, Ira G, Tercero JA, Holmes AM, Diffley JF, Haber JE (2004) Role of DNA replication proteins in double-strand break-induced recombination in Saccharomyces cerevisiae. Mol Cell Biol 24(16):6891–6899. https://doi.org/10.1128/mcb.24.16.6891-6899.2004

Xiao X, Lin WZ, Chou KC (2008) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29(12):2018–2024

Xu Y, Ding J, Wu L-Y, Chou K-C (2013) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8(2):e55844

Xu R, Lei L, Qin R, Huang Z, Chou K-C (2020) The topological entropy mechanism of coronavirus disease 2019 (COVID-19). Nat Sci 12(12):737–742

Yan C, Hu J, Wang Y (2008) Discrimination of outer membrane proteins using a K-nearest neighbor method. Amino Acids 35(1):65–73

Yang R, Zhang C, Gao R, Zhang L (2015) A machine learning approach to identify DNA replication proteins from sequence-derived features. 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE). IEEE, New York, pp 13–18

Yeeles JT, Janska A, Early A, Diffley JF (2017) How the eukaryotic replisome achieves rapid and efficient DNA replication. Mol Cell 65(1):105–116

Zhang R, Wei D-Q, Du Q-S, Chou K-C (2006) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2(3):309–314

Zhang C-J, Tang H, Li W-C, Lin H, Chen W, Chou K-C (2016) iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 7(43):69783

Zhang D, Fang L, Wang L, Pan Z, Lai Z, Wu M, Tang K, Ludan L, Qian D, Huang Z (2020) The chemical mechanism of pestilences or coronavirus disease 2019 (COVID-19). Nat Sci 12(11):717–725

Zheng H, Wei D-Q, Zhang R, Wang C, Wei H, Chou K-C (2007) Screening for new agonists against Alzheimer's disease. Med Chem 3(5):488–493

Zhou X-B, Chen C, Li Z-C, Zou X-Y (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248(3):546–551

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.