JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 30, Number 2, 2023

Pp. 204–222
DOI: 10.1089/cmb.2022.0241

Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models

AKSHARA PANDE,[1,*] SUMEET PATIYAL,[1,*] ANJALI LATHWAL,[1] CHAKIT ARORA,[1] DILRAJ KAUR,[1] ANJALI DHALL,[1] GAURAV MISHRA,[1,2] HARPREET KAUR,[1,3] NEELAM SHARMA,[1] SHIPRA JAIN,[1] SALMAN SADULLAH USMANI,[1,3] PIYUSH AGRAWAL,[1,3] RAJESH KUMAR,[1,3] VINOD KUMAR,[1,3] and GAJENDRA P.S. RAGHAVA[1]

ABSTRACT

In the last three decades, a wide range of protein features have been discovered to annotate a protein. Numerous attempts have been made to integrate these features in a software package/platform so that the user may compute a wide range of features from a single source. To complement the existing methods, we developed a method, Pfeature, for computing a wide range of protein features. Pfeature allows to compute more than 200,000 features required for predicting the overall function of a protein, residue-level annotation of a protein, and function of chemically modified peptides. It has six major modules, namely, composition, binary profiles, evolutionary information, structural features, patterns, and model building. Composition module facilitates to compute most of the existing compositional features, plus novel features. The binary profile of amino acid sequences allows to compute the fraction of each type of residue as well as its position. The evolutionary information module allows to compute evolutionary information of a protein in the form of a position-specific scoring matrix profile generated using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST); fit for annotation of a protein and its residues. A structural module was developed for computing of structural features/descriptors from a tertiary structure of a protein. These features are suitable to predict the therapeutic potential of a protein containing non-natural or chemically modified residues. The model-building module allows to implement various machine learning techniques for developing classification and regression models as well as feature selection. Pfeature also allows the generation of overlapping patterns and features from a protein. A user-friendly Pfeature is available as a web server python library and stand-alone package.

Keywords: binary profile, feature selection, machine learning techniques, protein composition, PSSM, Shannon entropy.

[1]Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
[2]Department of Electrical Engineering, Shiv Nadar University, Greater Noida, India.
[3]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India.
*These authors have contributed equally to this work.

1. INTRODUCTION

The advent of next generation sequencing technologies earmarked the production of a vast amount of genomic and proteomic data that lead to the exponential growth of protein databases. The experimental techniques used for functional annotation of a protein are time-consuming, expensive, and slow. Thus, the functional and structural annotation of a protein is one of the main tasks in bioinformatics. In the past, several hundred methods have been developed for predicting the functions of a protein that include the following: (1) overall function of a protein, (2) function of each residue, and (3) therapeutic annotation of chemically modified peptides. In all these methods, one of the significant challenges is to depict the numerical representation of a protein, known as protein feature or descriptor.

The last five decades have witnessed considerable progress in detecting new protein features. Functional annotation of a protein can be achieved by predicting the overall function of a protein, which includes subcellular localization, classification, and diversified properties of a protein. Subcellular localization of a protein is vital in understanding its function, such as TargetP (Emanuelsson et al, 2000), one of the first methods developed for subcellular localization.

In the last two decades, a wide range of methods have been developed for predicting subcellular localization of proteins in a cell that includes ESLpred (Bhasin and Raghava, 2004a), WoLF (Horton et al, 2007), PSLpred (Bhasin et al, 2005), HSLpred (Garg et al, 2005), RSLpred (Kaundal and Raghava, 2009), and MultiLoc2 (Blum et al, 2009). Similarly, methods has been developed for predicting proteins having a desired function or characteristic, which includes GPCRpred (Bhasin and Raghava, 2004b), NRpred (Bhasin and Raghava, 2004c), CytoPred (Lata and Raghava, 2008), and CyclinPred (Kalita et al, 2008).

Most of these methods utilize composition-based features to represent a variable length protein by a fixed length vector. In case of amino acid composition, a protein is represented by a vector of dimension 20. Numerous composition-based features such as pseudoamino acid, amphiphilic pseudoamino acid, auto-correlation, conjoint triad, and quasisequence order (QSO) have been reported over the years. Likewise, several methods have been developed to predict the therapeutic properties of a protein such as prediction of anticancer, antimicrobial, antibacterial, antifungal, antitubercular, and antihypertensive peptides (Agrawal et al, 2018; Gupta et al, 2013; Kumar et al, 2015; Lata et al, 2010; Manavalan et al, 2019; Manavalan et al, 2017; Meher et al, 2017; Sharma et al, 2013; Tyagi et al, 2013; Usmani et al, 2018a). Proteins with a therapeutic property play a vital role in designing protein-based drugs and vaccines (Dhall et al, 2020; Dhanda et al, 2017; Nagpal et al, 2017; Usmani et al, 2018b).

The last few decades witnessed a significant increase in the FDA approval of protein-based drugs, biomarkers, and vaccines (Usmani et al, 2017). In addition to the composition of a whole protein, the composition of N- or C-terminal residues of peptides has also been used for prediction (Lata et al, 2007). One of the major challenges in designing protein-based therapeutics is their stability in body fluids as the enzymatic system of the body degrades the protein. To minimize degradation of a protein, researchers are making chemical modifications in a protein. This is the reason most of the FDA-approved, protein-based drugs are chemically modified. Thus, computing descriptors of these chemically modified peptides is a challenge as amino acid sequence information is not adequate to represent these peptides.

Numerous methods have been developed in the past such as AntiMPmod (Agrawal and Raghava, 2018), CellPPDmod (Kumar et al, 2018), and HemoPI-MOD (Kumar et al, 2020) for predicting the therapeutic properties of chemically modified peptides. In these methods, structural descriptors are used for building prediction models; the structure of chemically modified peptides can be determined using protein modeling software such as PEPstrMOD (Singh et al, 2015).

Protein-level annotation provides overall property of a protein, but provides no information about its residues. Thus, there is need to annotate a protein at residue to understand the function of residues in a protein (Nagel et al, 2009). Chou-Fasman method (Chen et al, 2006) was one of the initial methods developed for predicting the secondary structure state of each residue in a protein. Similarly, methods have been developed for predicting the irregular secondary structure such as alpha turns (Wang et al, 2006), beta turns (Fuchs and Alix, 2005; Kaur and Raghava, 2003b; Kountouris and Hirst, 2010; Singh et al, 2015), gamma turns (Guruprasad and Rajkumar, 2000; Jahandideh et al, 2007), beta hairpins (de la Cruz et al, 2002), and beta barrel (Freeman and Wimley, 2012).

These methods allow residue-level annotation of a protein, where the function (secondary structure state) of each residue is computed. Binary profile is one of the widely used features for annotating a protein at residue level where a residue is represented by a vector of 20 dimensions. A pattern of length 17 residues

will be represented by a vector of dimension 340 (17×20). The performance of these methods improved significantly when evolutionary information was used instead of a single sequence (Rost and Sander, 1993). To utilize evolutionary information, protein profiles were generated to capture evolutionary information from similar sequences. One of the commonly used protein profiles is position-specific scoring matrix (PSSM) profile, which is generated using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (McGinnis and Madden, 2004). PSSM profiles have been heavily used as feature for predicting function of a protein at residue level.

In addition, methods have developed for predicting turns in proteins using a predicted secondary structure (Kaur and Raghava, 2003a; Kaur and Raghava, 2003b). In summary, there are four types of features: (1) compositional features, (2) binary profiles, (3) evolutionary profile, and (4) structural features.

Numerous web-based servers, stand-alone software, and libraries have been developed in the past that include PROFEAT (Li et al, 2006), PyBioMed (Dong et al, 2018), iFeature (Chen et al, 2018), protr (Xiao et al, 2015), Rcpi (Cao et al, 2015), propy (Cao et al, 2013b), PyDPI (Cao et al, 2013a), PDBparam (Nagarajan et al, 2016), POSSUM (Wang et al, 2017), and iFeatureOmega (Chen et al, 2022), for computing features of a protein. Despite a number of methods being developed in the past, still there are numerous important features that have not been integrated. Once we got the feature of a protein, the next challenge was to investigate the function of a protein.

One of the well-known methods for annotating a protein is similarity based search that assigns function based on sequence similarity; BLAST is one of the examples of such methods (Altschul et al, 1990). This technique fails in the absence of similarity between query sequence and sequences in the annotated databases. To overcome these limitations, a number of knowledge-based methods have been developed to predict the function of a protein. Most of the models developed in the past are based on machine learning techniques. Recently, a number of online platforms have been developed to allow users to build their own models using machine learning techniques; for example, iFeature, iLearn, BioSeq-Analysis, BioSeq-Analysis2.0, and iLearnPlus (Chen et al, 2020; Liu, 2019).

These online platforms allow many functions such as feature extraction, clustering, normalization, selection, dimensionality reduction, predictor construction, and best descriptor/model selection. These tools are designed for novice users who wish to develop prediction models from their data using optimal settings. The overall aim of these software/platform/libraries is to expedite annotation of proteins by involving a nontechnical researcher in building regression or classification models.

To supplement previous efforts, we have made a systematic attempt to develop a platform Pfeature that integrates most features discovered in the past along with the incorporation of some new features. In addition, it provides facility to implement machine learning techniques for building a prediction model on user's data. This platform provides six major modules (composition, binary, PSSM, structural, patterns, and models), following is a brief description of these modules. A composition-based module allows one to compute most of the features integrated in existing methods as well as novel features such as Shannon entropy, atom and bond composition, residue repeats, and distance distribution of residues. Binary profiles are generated based on amino acid, dipeptide, atom and bond, and amino acid index profile. PSSM module allows to compute various types of profiles based on evolutionary information. Structural module was developed to compute features of chemically modified proteins or peptides as modification cannot be presented by sequence. It allows to compute different types of features from a protein structure such as chemical fingerprints, secondary structure contents, and surface accessibility. Pattern-based features allow users to generate the patterns of desired window/pattern length from the existing features.

In addition, a module model has been developed for building prediction methods (classification or regression) using various machine learning techniques. This model building module includes many facilities such as merging features, clustering data, normalization of features, dimension reduction, feature selection, parameter optimization, cross-validation, and model construction. Pfeature will be available to the public in the form of a web server, stand-alone package, and python library.

2. MATERIALS AND METHODS

In this study, we mainly implemented existing algorithms for computing different types of features. In addition to the existing algorithms, we have also developed algorithms for specific type of features. Although these features have not been tested in any models, they do provide important information about a protein. In the following section, a summary of these algorithms has been briefly described.

2.1. Compositional features integrated in existing platforms

In the last 20 years, several composition-based features have been discovered for annotating a function or structure of proteins. Due to the discovery of composition-based features, researchers successfully used machine learning techniques in the field of bioinformatics for making protein classification or regression models. Composition-based features allow one to present proteins with varying lengths, by a fixed length vector. One of the simplest and commonly used composition-based features is amino acid or residue composition, where percent composition of 20 types of natural residues is computed in a protein. Thus, the amino acid composition of a protein of any length can be denoted by a vector of dimension 20. This concept has been successfully extended to dipeptide and tripeptide compositions, where 400 possible dipeptides and 8000 possible tripeptides are computed for a protein.

We know that all proteins are made up of 20 types of natural amino acid residues and they exhibit different types of physicochemical properties such as positive charge residues (e.g., lysine, arginine, histidine), polar residues (e.g., glutamine, asparagine, histidine), and hydrophobic residues (e.g., alanine, isoleucine, leucine). Thus, researchers also discovered and used features based on physicochemical properties, where composition of each property (or group of residues) is computed instead of each type of residue. Autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence. The amino acid properties used here are various types of amino acid indices (http://www.genome.ad.jp/dbget/aaindex.html). Three types of autocorrelation descriptors have been implemented in Pfeature, which are Normalized Moreau-Broto, Moran, and Geary autocorrelation (Ong et al, 2007).

In addition, there are a number of other commonly used features in literature that include the following: conjoint triad calculation, pseudoamino acid, composition-enhanced transition distribution, and many more. Almost all composition-based descriptors integrated in previous platforms have been integrated in our Pfeature to provide all features from a single platform.

2.2. Compositional features specific to Pfeature

There are a number of composition-based features that have been used in a number of studies successfully, but not integrated in any existing platforms. In addition, a number of new composition-based features have been described, which are not used in any study. Following is a brief description of these features.

2.2.1. Higher order dipeptide composition. In most of the existing platform, only traditional dipeptide ($i+1$) compositions have been integrated. In 2005, Garg et al introduced a new class of descriptors called higher order dipeptide compositions. It has been shown that hybrid models, Support Vector Machine (SVM) models, based on higher order dipeptides ($i+2$) perform better than models based on traditional dipeptide compositions. In the same study, it has been shown that the hybrid model that combines composition of different orders of dipeptide achieves maximum performance. Thus, we incorporate provision for computing any order of dipeptide in our platform. A detailed description of computing higher order dipeptides is available in figure 1 of Garg et al (2005). In brief, a traditional dipeptide ($i+1$) or a first-order dipeptide considers consecutive residues, for instance, traditional dipeptides for sequence ''AGRVMS'' will be AG, GR, RV, VM, and MS.

In case of second-order dipeptide ($i+2$), alternate residues will be considered, for example, for sequence ''ARGVMS,'' the dipeptides of second order will be AG, RV, GM, and VS. Similarly, dipeptides can be calculated for higher order such as third-, fourth-, and fifth-order dipeptides.

2.2.2. Amino acid repeats and distribution. In the past, a number of articles have shown the importance of repeats in proteins and their effect on protein structure and function (Faux, 2012; Kalita et al, 2006; Luo and Nijveen, 2014). Broadly, these repeats can be classified into the following two categories: (1) homorepeats responsible for several neurodegenerative and congenital malformation diseases, and (2) heterorepeats are diverse such as glycine–proline amino acid run in eukaryotic genomes, polar zippers, and asparagine-rich stretches (Perutz et al, 1994). There are a wide range of applications and functions of amino acid repeats in proteins. For example, for most of the cell penetrating peptides, there is a repeat of positively charged amino acids, particularly repeats of arginine (Agrawal et al, 2016; Perutz et al, 1994). Despite the importance of repeats, there is no method that captures the information of repeats in a protein sequence.

A simple composition of amino acids measures the fraction of residue, but not distribution of the repeats. Unfortunately, existing features do not measure the repeat of particular types of residues or their distribution. In this study, we have introduced new features that compute repeats of amino acids along with their distribution, as explained by Equations (1) and (2), respectively.

$$RRI_i = \frac{\sum_{j=1}^{N} (R_j)^2}{\sum_{j=1}^{N} R_j}, \tag{1}$$

$$DDOR_i = \frac{(R_{NT})^2 + \sum_{j=1}^{N} (R_j)^2 + (R_{CT})^2}{(L - F_i) + 1}, \tag{2}$$

where RRI_i and $DDOR_i$ are residue repeat information, and distance distribution of residue of type i, respectively. N and R_j are the maximum number of occurrences and number of runs/repeats of property type j, respectively. R_{NT}, R_j, R_{CT}, L, and F_i are residue distance from N-terminal, interdistance between residue type i, residue distance from C-terminal, total length of protein sequence, and frequency of residue type i, respectively.

2.2.3. Physicochemical property repeats. It has been observed in previous studies that certain types of residues having similar physicochemical properties (e.g., charged, polar, hydrophobic) are repeated in proteins; for example, membrane proteins have repeats of positively charged residues (Boyd and Beckwith, 1989). Thus, it is important to identify repeats of group of residues having a particular type of property. In this study, we have introduced new features that compute repeats of a particular property, as explained by Equation (3).

$$PRI_i = \frac{\sum_{j=1}^{N} (P_j)^2}{\sum_{j=1}^{N} P_j}, \tag{3}$$

where PRI_i, N, and P_j are property repeat information, maximum number of occurrences, and number of runs/repeats in occurrence j, respectively, for property type i.

2.2.4. Shannon entropy. In the field of information theory, entropy plays an important role to measure the information content. In the past, a number of articles have shown the importance of entropy in the field of protein informatics, following are a few examples. It has been shown that intrinsic disorder (fail to self-fold into fixed 3D structure) in proteins can be measured successfully by computing Shannon entropy of protein sequence (Romero et al, 2001). Shannon entropy indices were successfully used for predicting protein biomarkers in human colon cancer (Aguiar-Pulido et al, 2012). Recently, entropy is used for computing the structural feature of protein from its sequence (Bywater, 2015). It is clear from the above studies that entropy is an important descriptor that can be used to predict specific type of proteins.

To the best of our knowledge, there is no study that computes entropy-based features. To measure the level of complexity at the protein and residue levels, we compute Shannon entropy of a protein and entropy of each type of residues using Equations (4) and (5):

$$HS = -\sum_{i=1}^{20} p_i \log_2 p_i, \tag{4}$$

$$HR_i = -\, p_i \log_2 p_i, \tag{5}$$

where HS is Shannon entropy of a protein sequence and HR_i is entropy of a residue of type i. p_i is the probability of a given amino acid in the sequence. This equation was extended to compute entropy of a particular type of property such as charge, polarity, and hydrophobicity in a protein sequence.

2.2.5. Atom and bond composition. In case of amino acid composition, we compute the composition of each type of 20 residues in a protein. Each amino acid contains atoms and bond, these atoms and bond can be used as feature. Previously, this feature has been used for predicting subcellular location of proteins (Guralnik et al, 1991). Recently, this feature has been used to predict the cell penetrating potential of chemically modified peptides. It is difficult to present a chemically modified protein or peptide by a

simple amino acid or dipeptide composition as modified proteins also have chemical entities (Agrawal and Raghava, 2018; Mathur et al, 2018). Most of the FDA-approved, therapeutic peptides are chemically modified (Usmani et al, 2017).

Thus, this feature is important to predict the therapeutic potential of a peptide or protein. As far as we know, none of the existing platform allows to compute this feature. Thus, we incorporate this feature in Pfeature, where the user can calculate (1) atom, (2) bond, and (3) atom and bond composition.

2.2.6. PSSM composition. In the past, evolutionary information in the form of PSSM composition has been used for classification of proteins particularly for subcellular localization of proteins (Kumar et al, 2011; Kumar et al, 2007; Rashid et al, 2007). POSSUM is the only platform that allows to compute different types of descriptors based on PSSM profile. In PSSM profile, the evolutionary information is presented by a matrix of dimension $L \times 21$ (L rows and 21 columns) for a protein having length L. We have generated a vector of dimension 400 called PSSM-400, which maintains this composition (Kumar et al, 2007).

2.3. Binary profiles in existing studies

Composition-based features are suitable to develop models for classifying or predicting a protein with a desired function. However, these features are not sufficient to predict function at the residue level, such as predicting a secondary structure state of residues, adenosine triphosphate (ATP) interacting residues, and RNA binding residues. Qian and Sejnowski (1988) proposed a concept of binary profile for predicting a secondary structure of protein from its amino acid sequence. It has been shown in previous studies that the secondary structure of a residue not only depends on the residue but also depends on the neighbor residues. It means neighbor residues should be considered as input for predicting the secondary structure state of a particular residue.

Thus, Qian and Sejnowski (1988) use a window of length 13 for predicting a secondary structure of a query residue that includes six residues to the left, six residues to the right, and a query residue in the center. A major challenge was to convert a pattern/window of length 13 into numerical values as the input layer of a neural network. They represent each amino acid residue by a vector of length 21, where 20 U for 20 natural amino acids and 1 U for adding dummy amino acid at both termini to create one pattern corresponding to each number. Each amino acid is presented by a vector of dimension 21; if amino acid is not present, then value of element is 0, otherwise one. Thus, a pattern of length 13 will be represented by a vector of dimension 13×21.

After this study, binary profiles have been widely used for residue-level annotation that includes prediction of protein's secondary structure as well as nucleotide or ligand binding sites in a protein (Ansari and Raghava, 2010b; Kaur and Raghava, 2004a; Kaur and Raghava, 2004b; Kaur and Raghava, 2003a; Kaur and Raghava, 2003b; Kumar et al, 2007; Panwar et al, 2013; Patiyal et al, 2020). We have integrated binary profile in Pfeature, the methodology has been adopted from previous studies (Chauhan et al, 2013; Chauhan et al, 2012; Singh et al, 2015).

2.4. Novel binary profiles

In addition to a standard binary profile that is heavily used for annotating protein, we introduce some new features. These features are based on our experience and logics; we hope that these new features will be useful to the researchers working in the area of functional annotation. Following is a brief description of profiles integrated in our platform.

2.4.1. Dipeptide profile. It has been shown in a number of studies that a dipeptide composition provides more information than a simple amino acid composition. Thus, a model based on dipeptide composition performs better than a model based on amino acid composition. In case of dipeptide, composition of residues is further divided into 20 compositions based on the type of next neighbor residue. It means that a dipeptide composition provides more information due to information of the adjacent residue. In the past, amino acid has been used to annotate residues based on a binary or amino acid profile. This inspires us to introduce a new feature, dipeptide profile, to annotate a protein at residue level.

We have created a vector of dimension 400 for 20 amino acids and each dipeptide is represented by a vector of dimension 400, where the presence of dipeptide is represented by one and the absence by zero. This feature has not been tested so far in any prediction method, but we believe it can be of great use in future studies.

2.4.2. AAindex profile. AAindex is a database of amino acid indices where each AAindex is a set of 20 numerical values representing various physicochemical and biochemical properties of amino acids (Kawashima and Kanehisa, 2000; Kawashima et al, 2008; Kawashima et al, 1999). The current version of database maintains 566 amino acid indices. These indices are heavily used in literature as a feature for building classification methods. Most of the existing platforms allow to compute composition of indices. To the best of our knowledge none of the methods or platforms uses AAindex-based profile. Each of the indices provides highly valuable information, and thus, we proposed the AAindex-based binary profile in this study.

In this study, we have used AAindex values for computing binary profiles. In this method, we have represented each amino acid residue with a vector of dimension 553, and hence, if a protein sequence has L residues, the resulting vector will be of $553 \times L$ size. This function gives the binary profile of protein based on AA indices. If normalized score of AAIndex value of a particular residue is negative, the function assigns ''0'' to that residue, otherwise assigns ''1.'' For example, the binary profile for amino acid ''A'' for the following AAindex values (ARGP820101, ANDN920101, ARGP820102, ARGP820103, BEGF750101, BEGF750102, BEGF750103, BHAR880101, BIGC670101, and BIOV880101) will be {0, 0, 1, 1, 1, 1, 0, 0, 0, 0}.

2.4.3. Property profile. Physicochemical properties of amino acids have been used heavily in the field of protein annotation. Each type of 20 residues exhibits a different property that includes positive charge, polar, hydrophobicity, hydrophilicity, aromaticity, aliphatic, and so on. In most of the studies, these properties are used to compute the composition of residues based on their physicochemical properties. In 2010, physicochemical properties have been used for prediction of conformational epitope (Ansari and Raghava, 2010a). In this study, we have represented a protein by the physicochemical properties of its residues. Here, we have represented each residue with the vector size of 25, where each element represents a particular physicochemical property. We present an amino acid by ''1'' if it is positively charged, otherwise ''0.'' This way, we may create property profiles for 25 types of physicochemical properties such as hydrophobicity, hydrophilicity, and polarity.

2.5. Evolutionary information in the form of PSSM profile

The evolutionary information of protein brings evolution in the field of protein secondary structure prediction (Rost and Sander, 1993). The performance of models based on evolutionary information is significantly better than the methods based on sequence. To generate evolutionary information for a protein, the following process is performed: (1) search query protein sequence in databases to get similar protein sequences, (2) these similar sequences are aligned using multiple sequence alignment software, (3) representative profile is generated from aligned sequences, (4) these representative profiles are searched in databases to extract remote homologous sequences, and (5) again sequences are aligned using step 2, and profile is generated using step 3. This process is repeated to get an optimized profile.

PSI-BLAST is a commonly used software that integrates all the required steps to generate a protein profile for a given protein. It generates profile in the form of PSSM profile. In the past, most of the studies used PSSM profile to annotate function of protein at residue level (Kumar et al, 2008). To provide service to users based on PSSM profile, we integrate facility to compute PSSM profile. In this study, evolutionary information is extracted in the form of PSSM profiles, generated by PSI-BLAST. Pfeature allows one to generate PSSM-based features called PSSM profile. These PSSM profiles are used to annotate the protein at residue level.

Numerous studies demonstrate that PSSM composition-based models perform better than amino acid composition in the classification of proteins (Kaur et al, 2020; Kumar et al, 2007; Verma et al, 2010). Thus, we integrate facility to generate PSSM composition for a given protein. In the past, a number of techniques have been used to normalize PSSM; in this study, we used the following techniques for the sake of normalization.

- **pssm_n1**: Due to the large number of variations in the value of PSSM, it is necessary to normalize it. In this method, each element of matrix is normalized by formula $\frac{1}{(1+e^{-x})}$, where x is the element.
- **pssm_n2**: This is the second technique to normalize the elements of PSSM using the formula $\frac{(x-\min)}{(\max-\min)}$, where x is the element, and min and max represent the minimum and maximum value, respectively.

- **pssm_n3**: This is the third technique to normalize the matrix using the formula $\frac{(x-\min)}{(\max-\min)} \times 100$, where x is the element, and min and max represent the minimum and maximum value, respectively.
- **pssm_n4**: This is the fourth technique to normalize the PSSM profile using the formula $\frac{1}{(1+e^{-x/100})}$, where x is the element.

The PSSM contained the probability of occurrence of each type of amino acid residue at each position along with insertion/deletion. It measures conservation of a residue in a given location in protein. This means that evolutionary information for each amino acid is encapsulated in a vector of 21 dimensions, where the size of PSSM of a protein with N residues is NX21. For example, PSSM profile of a pattern/peptide of 9 amino acids will be a dimension of 9X21.

2.6. Structural features for chemically modified proteins

As described above, composition-based feature allows one to annotate the overall function of protein, and binary profile allows to annotate protein at the residue level. Evolutionary information permits to improve the performance of a prediction model. All these features are designed for interpreting proteins/peptides that contain only natural amino acids. These features are also called sequence-based features, as they are generated from primary sequence of a protein. None of these features is applicable for developing models for chemically modified proteins that contain non-natural amino acids or chemically modified residues. It has been observed that most of the FDA-approved, protein-based drugs are chemically modified, as natural proteins have a number of limitations such as fast degradation and toxicity (Usmani et al, 2017). It means these features are not suitable for designing therapeutic proteins or predicting the therapeutic potential of a protein. In addition, these sequence-based features are not adequate for structural annotation of proteins.

In summary, sequence-based features are not suitable for predicting the structure or therapeutic potential of a protein. To facilitate the community for designing therapeutic proteins or structural annotation of a protein, we incorporate structural features in Pfeature. As chemical modifications in proteins or structure of protein cannot be represented by sequence, input of module is the protein structure. Following is a brief description of structural descriptors incorporated in this study.

2.6.1. Fingerprints. One of the most successful chemical descriptors in cheminformatics are fingerprints, which are heavily used for functional annotation of chemicals (Dhanda et al, 2013; Singh et al, 2015; Singla et al, 2013). In the recent articles, application of fingerprints has been extended to predict the therapeutic potential of chemically modified peptides (Agrawal and Raghava, 2018; Kumar et al, 2020; Kumar et al, 2018). To generate fingerprints of peptides/proteins, this module needs the tertiary structure of protein. One can predict the tertiary structure of a chemically modified peptide using state-of-the art software PEPstrMOD (Singh et al, 2015). This module of Pfeature was developed to calculate different types of fingerprints from tertiary structure of protein. The fingerprints were calculated using PaDEL software (Yap, 2011), which is a java-based software.

In this module, PaDEL software provides 10 different fingerprint types, which in total provide 14,532 fingerprint values. These fingerprints are calculated using mainly The Chemistry Development Kit (CDK). Along with CDK, other fingerprints are also computed such as PubChem fingerprints, molecular access system (MACCS) fingerprints, and Klekota-Roth fingerprints.

2.6.2. Simplified Molecular Input Line Entry System. SMILES stand for Simplified Molecular Input Line Entry System. It is a type of line notation for representing various molecules and reactions. It contains the same information as the extended data tables. One of the advantages of using SMILES is, it easy to understand since it is a linguistic construct rather than a computer data structure. Also, the SMILES format takes 50%–70% less space in comparison with other ways of representing the information as well as it required lesser time for processing the information. SMILES notation is represented by a series of characters, and no spaces are present in between the characters. SMILES notation follows five simple rules required for its encoding, which are corresponding to atoms, bonds, branches, ring closures, and disconnections. A detailed description of the SMILES notations can be obtained at http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

3. RESULTS

The major aim of Pfeature is to facilitate the users in computing most of existing features discovered in the literature, along with some new features. It can be used as web server, where users can submit their sequence/structure on our website for computing features. In addition, python scripts, stand-alone, and libraries have been developed, so that the user can compute features on a local machine. Following is a brief description of different types of implementation and facilities provided by Pfeature.

3.1. Web implementation

A web server has been developed on Linux/Ubuntu operating system using Apache software. Web pages of this server have been developed using HTML, PHP5, and CSS3. Responsive web design has been used to provide compatibility with wide ranges of devices (e.g., iPad, smartphone, laptop, desktop). Submission page allows users to submit protein sequences in FASTA format or in single line. The users are also allowed to provide PDB IDs or UniProt IDs as input. In addition to displaying results as HTML page, the server allows to download results in csv format. Major modules of Pfeature are shown in Figure 1; following is a brief description of menus/submenus.

3.1.1. Composition menu. This module is designed for computing a wide range of composition-based features. These features have been classified in the following five submenus. This module allows user to compute simple composition-based features such as amino acid composition, dipeptide composition, and tripeptide composition. This server also allows to compute atom and bond composition of a protein (Guralnik et al, 1991). These features can be used for predicting function of a chemically modified protein (Agrawal and Raghava, 2018; Mathur et al, 2018). In case of dipeptide composition, a server allows to compute traditional as well as higher order dipeptide composition (Garg et al, 2005). In case of standard physicochemical properties, a server computes composition of around 19 types of physicochemical properties. Submenu AAindex allows user to compute composition of any amino acid index out of a total of 566 aa indices present in database AAindex version 9.0.

Repeats and distribution are an interesting module that allows to calculate novel features such as repeats and distribution of residues in a protein. In the past, it has been shown that entropy can be used to predict intrinsic disorder in a protein (Romero et al, 2001). Thus, server also allows to compute Shannon entropy of a protein as well as each type of residue in a protein.

Submenu miscellaneous allows user to compute a wide range of composition-based features described in previous studies and platforms (Dong et al, 2018; Li et al, 2006). This submenu has seven options to compute complex composition-based features; following is a brief description of each feature. Autocorrelation computes distribution of amino acid properties in the protein/peptide sequences. In case of

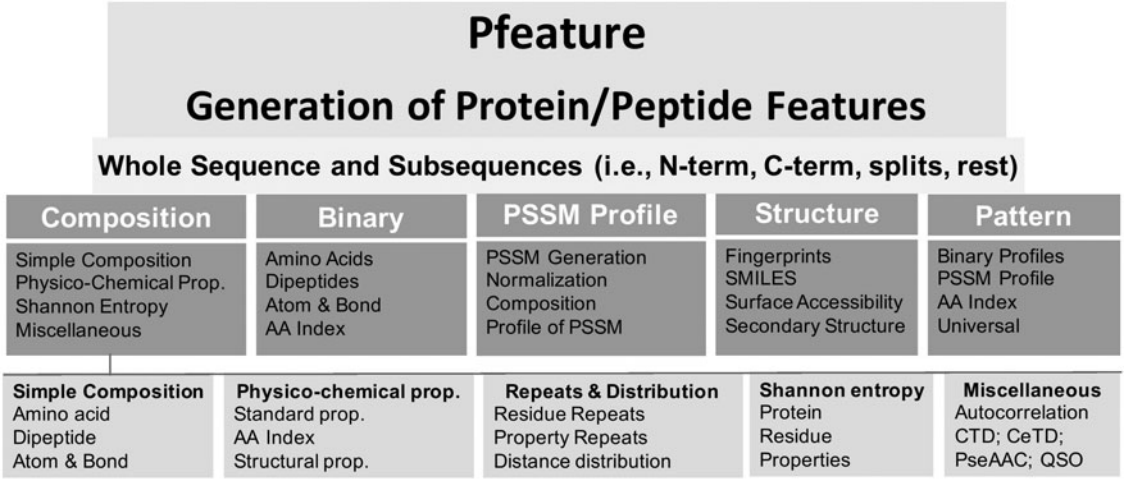

FIG. 1. Major menus and submenus in Pfeature for computing protein features.

conjoint triad descriptors, amino acids are divided into seven groups, then properties of triad are computed, and output vectors have a dimension of 343 ($7\times7\times7$). Composition enhanced transition, and distribution computes enhanced transition of residues after grouping the amino acid residues into three groups based on seven attributes. Pseudo- and amphiphilic pseudoamino acid compositions are one of the most widely used protein features implemented by a scientific community (Chou, 2005; Chou, 2001).

QSO descriptors are computed from the distance matrix between the 20 amino acids. Sequence-order coupling number allows user to compute a set of sequence-order-coupling numbers (Chou, 2000).

3.1.2. Binary profiles. The composition-based features described in the above menu are not suitable to predict function of residues in a protein, such as predicting protein secondary structure (Singh et al, 2015; Wang et al, 2006). To facilitate users for developing methods for predicting function of residues in a protein, we developed a module for generating binary profiles. Binary profile menu has a number of submenus that allow to compute different types of binary profiles. Amino acid submenu permits users to generate the binary profiles corresponding to each residue of the peptide or protein sequences. Dipeptide submenu calculates the dipeptide binary profile with sn option to select the order of dipeptide.

Submenu atom and bond is subdivided into three sections as atom, bond, and atom and bond, generating the binary profiles for atoms, bonds, and both, respectively. In case of atoms, a vector will be represented by five values, based on five types of atom (C, H, N, O, and S). In case of bonds, vector dimension is of length four, based on four types of bond (cyclic, benzene ring, single bond, and double bond). Residue property submenu generates binary profiles for each residue in the amino acid sequence, which apprehends the particular physicochemical properties for each input sequence. It generated the output vector of length $25\times L$, where L is the protein/peptide length. AAIndex submenu allows the user to compute the binary profiles corresponding to 553 aaindices.

3.1.3. Evolutionary information. The evolutionary information of protein brings evolution in the field of protein secondary structure prediction (Rost and Sander, 1993). The performance of models based on evolutionary information is significantly better than methods based on a single sequence (Kaur et al, 2020; Kumar et al, 2007; Verma et al, 2010). In most of these studies, the PSSM profile is computed to capture the evolutionary information of a protein. To compute different types of PSSM-based compositions, menu evolutionary information has been integrated. This menu captures the evolutionary information, which plays a significant role in defining the functionality of the peptide or protein. It has a number of submenus for generating PSSM profile for a protein sequence, where PSSM profile is generated using PSI-BLAST. This submenu captures the spatial information for evolutionary conservation of amino acids. Submenu normalization of PSSM allows user to generate a normalized PSSM.

In this server, commonly used four types of normalization methods have been implemented. Composition of PSSM allows user to calculate the composition of PSSM. This method gives output in the form of 400 values. Submenu profile of PSSM is developed for computing a normalized PSSM. This method gives output in the form of matrix of $N\times21$ for the peptide or protein sequence of length N.

3.1.4. Structural descriptors. The composition-based feature allows one to annotate the overall function of a protein, and binary profile allows one to annotate a protein at residue level, whereas evolutionary information allows to improve the performance prediction models. All these features are designed for interpreting proteins that contain only natural amino acids. None of these features is applicable for developing methods for chemically modified proteins that contain non-natural amino acids or chemically modified residues. It has been observed that most of the FDA-approved, protein-based drugs are chemically modified, as natural proteins have a number of limitations such as fast degradation (Usmani et al, 2017). This menu is designed for computing structural descriptors from tertiary structure of a protein, which can be used for predicting functions of chemically modified proteins.

Submenu fingerprint is designed to calculate the fingerprint from tertiary structure of a protein; it needs PDB file or PDB ID as input. These fingerprints are calculated using PaDEL software (Yap, 2011). SMILES submenu provides the SMILES notation from structure. SMILES is generated using Open Babel software and provides a line notation for representing peptide or protein structure. Surface accessibility submenu is built to calculate the relative surface accessibility of the structure, using NACCESS software. The output exhibits the surface accessibility for each residue in a protein sequence. Secondary structure submenu allows to assign the secondary structure of a protein from its tertiary structure and calculates the alpha helix, sheet, and coil composition, implemented using DSSP software (Touw et al, 2015).

3.1.5. Menu for model building. All the menus described above allow researchers to compute features of a protein. These features can be used for building the models for annotation of a protein or its residue (Bhasin and Raghava, 2004a; Buchan and Jones, 2019; Garg et al, 2005). These models are developed by an individual researcher having expertise in building models and in feature engineering. Recently, a number of user-friendly platforms have been developed that allow users to build a model just by uploading their data (Chen et al, 2020; Liu, 2019; Liu et al, 2019). To complement, we add a module called building models in Pfeature, which incorporate a number of facilities required for building models including feature engineering.

The model building menu of Pfeature has a number of submenus that include the following: (1) merging features, (2) selection of relevant features, (3) classification, and (4) prediction. Our feature merging option allows one combined feature from two files. Another option in this module is to identify a relevant feature as all features are not equally important. These options allow one to perform feature engineering-related tasks. A major feature of this module is to build classification or prediction models. Users may either paste or upload the training and independent data sets separately, formatted as csv files on the web interface of the server. Input files must have the last column as labels, used while building classification models. Various selectable submodules are provided, if user wants to implement normalization, feature selection, dimensionality reduction, and clustering on the data set.

Different machine learning algorithms are implemented with additional functionalities such as parameter optimization and K-fold cross-validation. The output is downloadable in zip file containing all the results on training and independent data sets.

3.1.6. Substring of a sequence. In most of the protein classification methods, features are computed from a whole protein. In literature, it has been shown that the split amino acid composition (SAAP)-based method performs better than the whole-sequence composition-based method (Restrepo-Montoya et al, 2009). In SAAP, proteins are split in a number of parts and then the composition of each part is computed. It has been observed in the past that N-terminal of a protein is also responsible for its function, for example, classical secretory proteins containing signal peptide. A short peptide presents at the N-terminus of the majority of proteins that are destined toward the secretory pathway. The most common endoplasmic reticulum retention signal is the amino acid sequence KDEL or HDEL at the C-terminus. This keeps the protein in the endoplasmic reticulum and prevents it from entering the secretory pathway.

Pfeature allows user to compute a wide range of features in selected regions of a protein such as terminal (C- or N- terminal), rest, and split of a protein. One of the advantages of selecting a region is that user can generate both composition and binary profile for this fixed length selected region (Lata et al, 2007).

3.2. Stand-alone version and library

In addition to web-based implementation, we have also developed stand-alone of Pfeature, suitable to run on local machines. It contains more than 120 functions written in python and allows to compute all features of a protein. It is written in Python3; hence, it can be run on any operating system containing Python3. To develop a method for annotating proteins in python, user may wish to call python functions for speed optimization. We have also developed python library of Pfeature, which can be downloaded from https://github.com/raghavagps/Pfeature. The prerequisite to run the python library is pandas, NumPy, and python version 3.6 or above. The command line usage of the stand-alone version is shown in Supplementary Figures S1–S3.

3.3. Comparison with existing platforms

In the past, more than 20 web servers/platforms/stand-alone software packages have already been developed, why another platform for computing protein features? Thus, it is important to justify the newly developed method Pfeature, which is developed to complement the existing methods. One to one comparison of existing methods with Pfeature is not feasible as a number of existing methods also allow to compute DNA, RNA, chemical, and protein features, whereas Pfeature is only for the peptides or proteins. Composition-based features integrated in Pfeature are shown in Table 1, it also shows existing methods that integrate these features. We have divided exiting methods into two categories: feature calculation and analysis/prediction methods. Although one can use analysis and prediction methods (e.g., iFeature, iLearn, iLearnPlus, BioSeq-Analysis, BioSeq-Analysis2.0) for calculating features, they are not user friendly.

Table 1. Comparison of Composition-Based Features Integrated in Different Platforms/Software

Features integrated in Pfeature		*Feature calculation software*	*Analysis and prediction platform*
Type of descriptors	*No.*		
Amino acid	20	All[a]	All
Dipeptide	400	All	All
Tripeptide	8000	All	All
Atom and bond	9	None	None
Physicochemical (standard + advanced)	30	Most	Most
AAIndex	566	Most	Most
Autocorrelation (3 × AAIndex)	1698	Most	Most
Shannon entropy (overall + each residue)	21	None	None
Shannon entropy of property	24	None	None
Distance distribution	20	None	None
Repeats of residues	20	None	None
Repeats of physicochemical properties	19	None	None
Pseudoamino acid	$20 + \lambda$	Most	BSA2, iLP
Amphiphilic pseudoamino acid	$20 + \lambda \times 3$	Most	BSA2, iLP
Conjoint triad calculation	343	Most	BSA2, iLP
CETD	189	Most	Most
SOCN	$\lambda \times 2$	Most	BSA2, ILP
QSO	$40 + (\lambda \times 2)$	Most	BSA2, iLP
PSSM composition	400	POSSUM	BSA2
No. of descriptors for whole protein (only single value of $\lambda = 5$)			11,879
Total descriptors (whole protein + N-term + C-term + RN-Term + Split etc.)			95,137

These descriptors are computed at protein level, and can be used to compute the overall function/structure of a protein.

[a]All: All software; Most: Most of the software; None: Only Pfeature; BSA2: BioSeq-Analysis 2.0; iLP: iLearnPlus.

CETD, composition-enhanced transition distribution; PSSM, position-specific scoring matrix; QSO, quasisequence order; SOCN, sequence order coupling number.

As shown in Table 1, a large number of features only integrated in Pfeature such as Shannon entropy, distance distribution, and repeats. Although POSSUM allows to compute a wide range of PSSM profiles, it does not allow to compute other features. In addition, Pfeature allows to compute the composition of different regions/segments of a protein (such as N-terminal residues, C-terminal residues, split) which have been used successfully in a number of studies. To the best of our knowledge, this feature is not incorporated in any platform.

One of the unique characters of Pfeature is that it allows to compute features at the residue and atom level, which is required to annotate residues of a protein (see Table 2). This table also shows existing methods that support these features. As shown in Table 2, most of the feature calculation software does not support these features except POSSUM, which only supports PSSM profile. Analysis and prediction method BioSeq-Analysis2.0 supports few features and iLearnPlus supports only one feature. This clearly demonstrates that Pfeature allows to compute many important features that are not integrated in any existing software.

As shown in Table 3, Pfeature is available as a web server, stand-alone package, and python library. It also allows to compute features directly rather than digging features from analysis and prediction. It is important to mention here that Pfeature is a user-friendly web server where a specific web page is designed for computing each type of a feature separately. Most of the existing web servers have used the same template (single page) for computing features or analysis and prediction, such as the web page of servers such as BioSeq-Analysis, BioSeq-Analysis, iLearnPlus, iLearn, POSSUM, Pse-in-One, and Pse-in-One 2.0. Pfeature provides diversity in web design, and it provides a wide range of flexibility to users for computing any desired feature.

3.4. Case study: Applications of Pfeature

In the past, various studies have shown the importance of structure-based features for the prediction of therapeutic potential of chemically modified proteins (Agrawal and Raghava, 2018; Kumar et al, 2020).

Table 2. Comparison of Residue- and Atom-Level Features Integrated in Different Platforms/Software

Features integrated in Pfeature		*Feature calculation software*	*Analysis and prediction platform*
Type of descriptors	*No.*		
Binary profiles			
Amino acid	$L\times20$	None[a]	BSA2, iLP
Dipeptides	$L\times400$	None[a]	None
Physicochemical Properties	$L\times25$	None[a]	None
AAIndex	$L\times566$	None[a]	BSA2
Atom + bond	$L\times20\times9$	None[a]	None
PSSM profiles			
PSSM raw profile	$L\times21$	POSSUM	BSA2
normalized pssm profile	$L\times21$	POSSUM	BSA2
Structural descriptors			
Fingerprints	14,532	None[a]	None
Similes format	$L\times15$	None[a]	None
Surface accessibility	L	None[a]	BSA2
Average secondary Structure	3	None[a]	BSA2
Total descriptors, if we take protein length ($L=100$)			139,435

These descriptors are suitable for predicting function of residues in a protein and function of chemically modified proteins.
[a]None: Only Pfeature; BSA2: BioSeqAnalysis 2.0; iLP: iLearnPlus.

Pfeature also provides the facility to the users to compute the structure-based features/descriptors (such as SMILES and fingerprints) using the ''Structure'' module, which can be used for the prediction of function of chemically modified proteins/peptides. Sofi and ArifWani (2021) developed a computational model for the prediction of amyloid proteins using structure information where they utilized Pfeature ''Structure'' module to compute the secondary structures and solvent accessibility-based features. Shahraki et al (2022) also developed a computational approach for targeted discovery of biocatalysts from metagenomic data where they have used atom and bond compositions, Shannon entropy, residue repeats, and distance distribution of residue-based novel features of Pfeature tool.

Hassan et al (2020) measure flow of SARS-COV-2 based on ACE2 receptor features computed using novel features of Pfeature. In addition, several recent studies such as IL6pred, IL13pred, AlgPred2.0,

Table 3. Comparison of Features of Different Software/Platforms from User Point of View

Software/platform	*Year*	*Web server*	*Stand-alone*	*Library*	*Features*[a]	*Prediction*
Pfeature	2022	Yes	Yes	Python	Direct	Yes
iFeatureOmega	2022	Yes	Yes	Python	Indirect	Yes
iLearnPlus	2021	Yes	Yes	No	Indirect	Yes
iLearn	2019	Yes	Yes	No	Indirect	Yes
PyFeat	2019	No	Yes	No	Direct	Yes
BioSeq-Analysis2.0	2019	Yes	Yes	No	Indirect	Yes
iFeature	2018	Yes	Yes	No	Indirect	Yes
PyBioMed	2018	No	Yes	Python	Direct	No
POSSUM	2017	Yes	Yes	No	Direct	No
BioSeq-Analysis	2017	Yes	Yes	No	Indirect	Yes
Pse-in-One 2.0	2017	Yes	Yes	No	Direct	No
PDBparam	2016	Yes	No	No	Direct	No
BioTriangle	2016	Yes	No	No	Direct	No
Pse-in-One	2015	Yes	Yes	No	Direct	No
Protr/ProtrWeb	2015	yes	Yes	R	Direct	No
PyDPI	2013	No	Yes	Python	Direct	No
Propy	2013	No	Yes	No	Direct	No
PROFEAT	2011	Not active	No	No	Direct	No

[a]Direct: Developed specifically for calculating features; Indirect: Developed for building prediction models where features can be extracted from the results.

B3Pred, HLAncPred, and ABCRpred (Dhall et al, 2022; Dhall et al, 2021; Jain et al, 2022; Kumar et al, 2021; Maryam et al, 2021; Sharma et al, 2020) used composition-based module and novel features of Pfeature to calculate the sequence-based features to develop prediction models.

4. DISCUSSION

PROFEAT is one of the first web servers developed in 2006, for computing structural and physicochemical features of proteins from its amino acid sequence (Li et al, 2006). Its updated versions were published in 2011 that incorporate network, segment descriptors, and topological descriptors. A large number of features in PROFEAT such as network descriptors are not integrated in Pfeature. Similarly, a python library PyBioMed has been developed for computing a wide range of features of DNA, protein, and chemical molecules (Dong et al, 2018). PyDPI is feature-rich, python-based stand-alone package that computes 52 types of protein features from 6 feature groups (Cao et al, 2013a).

Recently, one of the powerful packages, iFeature, has been developed that compute a comprehensive spectrum of 18 major sequence encoding schemes that encompass 53 different types of feature descriptors. It also integrates 12 different types of commonly used feature clustering, selection, and dimensionality reduction algorithms. These features are important for developing machine learning technique-based models for predicting the function of biomolecules. Recently, updates of iFeature have been developed called iLearn that integrates more features. In this series, iLearnPlus is the most powerful platform for analysis and prediction. Similarly, BioSeq-analysis2.0 has been developed, which is an update of BioSeq-analysis2.0. Both iLearnPlus and BioSeq-analysis2.0 integrate a wide range of features and allow to predict/analyze protein data.

In summary, a number of software packages have been developed to compute descriptors of biological and chemical molecules. Each package has a unique set of features that complement other existing packages. The aim of developing Pfeature is to complement the existing platform so that user may get more options in the field of bioinformatics. We are particularly interested in developing features for computing the function of each residue in a protein. None of the existing servers provides facility to compute features for annotating chemically modified proteins and annotation of protein. In addition, existing servers do not have the facility to compute composition of a specific region of protein. To overcome some of the limitations of existing servers, we integrate most of the descriptors described in literature as well as some of the novel descriptors in Pfeature.

In summary, it is a comprehensive, easy-to-use python package that computes a large number of features, and allows users to calculate various features of proteins/peptides based on their sequence, structure, and physiochemical properties. Recently, features computed using Pfeature have been used for developing models for predicting pattern recognition receptors, interleukin-6 inducing peptides, and allergenic peptides (Dhall et al, 2020; Kaur et al, 2020; Sharma et al, 2020). We have also provided the model building module that can be used independently to analyze the descriptors and build different machine learning-based models for classification and regression.

AUTHORS' CONTRIBUTIONS

S.P., A.P., A.L., C.A., D.K., A.D., S.J., and G.M. wrote all the scripts. H.K., S.P., A.D., A.L., C.A., D.K., S.S.U., P.A., R.K., S.J., and V.K. developed the web interface. N.S., S.J., A.L., and C.A. prepared the manual. S.S.U., S.P., A.P., N.S., R.K., and A.D. prepared the first draft of the article. S.S.U. and G.P.S.R. prepared the final version of the article. G.P.S.R. conceived the idea and coordinated the entire project.

ACKNOWLEDGMENTS

The authors are thankful to the funding agencies Department of Biotechnology (DBT) and Department of Science and Technology (DST-INSPIRE) and Council of Scientific and Industrial Research (CSIR), Government of India, for financial support and fellowships. BioRxiv link: https://www.biorxiv.org/content/10.1101/599126v1

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This work was supported by the Department of Biotechnology (DBT), Government of India, India. Grant Number: BT/PR40158/BTIS/137/24/2021.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3

REFERENCES

Agrawal P, Bhalla S, Chaudhary K, et al. In silico approach for prediction of antifungal peptides. Front Microbiol 2018;9:323; doi: 10.3389/fmicb.2018.00323

Agrawal P, Bhalla S, Usmani SS, et al. CPPsite 2.0: A repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res 2016;44(D1):D1098–D1103; doi: 10.1093/nar/gkv1266

Agrawal P, Raghava GPS. Prediction of antimicrobial potential of a chemically modified peptide from its tertiary structure. Front Microbiol 2018;9:2551; doi: 10.3389/fmicb.2018.02551

Aguiar-Pulido V, Munteanu CR, Seoane JA, et al. Naive Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer. Mol BioSyst 2012;8(6):1716–1722; doi: 10.1039/c2mb25039j

Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol 1990;215(3):403–410; doi: 10.1016/S0022-2836(05)80360-2

Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res 2010a;6(1):6; doi: 10.1186/1745-7580-6-6

Ansari HR, Raghava GPS. Identification of NAD interacting residues in proteins. BMC Bioinformatics 2010b;11(1):160; doi: 10.1186/1471-2105-11-160

Bhasin M, Garg A, Raghava GPS. PSLpred: Prediction of subcellular localization of bacterial proteins. Bioinformatics 2005;21(10):2522–2524; doi: 10.1093/bioinformatics/bti309

Bhasin M, Raghava GPS. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 2004a;279(22):23262–23266; doi: 10.1074/jbc.M401932200

Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 2004b;32(Web Server):W414–W419; doi: 10.1093/nar/gkh350

Bhasin M,d Raghava GPS. GPCRpred: An SVM-based method for prediction of families and subfamilies of g-protein coupled receptors. Nucleic Acids Res 2004c;32(Web Server):W383–W389; doi: 10.1093/nar/gkh416

Blum T, Briesemeister S, Kohlbacher O. MultiLoc2: Integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 2009;10(1):274; doi: 10.1186/1471-2105-10-274

Boyd D, Beckwith J. Positively charged amino acid residues can act as topogenic determinants in membrane proteins. Proc Natl Acad Sci U S A 1989;86(23):9446–9450; doi: 10.1073/pnas.86.23.9446

Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 Years on. Nucleic Acids Res 2019;47(W1):W402–W407; doi: 10.1093/nar/gkz297

Bywater RP. Prediction of protein structural features from sequence data based on Shannon entropy and Kolmogorov complexity. PLoS One 2015;10(4):e0119306; doi: 10.1371/journal.pone.0119306

Cao D-S, Liang Y-Z, Yan J, et al. PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. J Chem Inf Model 2013a;53(11):3086–3096; doi: 10.1021/ci400127q

Cao D-S, Xiao N, Xu Q-S, et al. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics 2015;31(2):279–281; doi: 10.1093/bioinformatics/btu624

Cao D-S, Xu Q-S, Liang Y-Z. Propy: A tool to generate various modes of Chou's PseAAC. Bioinformatics 2013b;29(7):960–962; doi: 10.1093/bioinformatics/btt072

Chauhan JS, Bhat AH, Raghava GPS, et al. GlycoPP: A webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. PLoS One 2012;7(7):e40155; doi: 10.1371/journal.pone.0040155

Chauhan JS, Rao A, Raghava GPS. In silico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences. PLoS One 2013;8(6):e67008; doi: 10.1371/journal.pone.0067008

Chen H, Gu F, Huang Z. Improved Chou-Fasman method for protein secondary structure prediction. BMC Bioinformatics 2006;7 Suppl 4:S14; doi: 10.1186/1471-2105-7-S4-S14

Chen Z, Liu X, Zhao P, et al. IFeatureOmega: An integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. Nucleic Acids Res 2022(W1):W434–W447; doi: 10.1093/nar/gkac351

Chen Z, Zhao P, Li F, et al. iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34(14):2499–2502; doi: 10.1093/bioinformatics/bty140

Chen Z, Zhao P, Li F, et al. iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform 2020;21(3):1047–1057; doi: 10.1093/bib/bbz041

Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 2000;278(2):477–483; doi: 10.1006/bbrc.2000.3815

Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43(3):246–255; doi: 10.1002/prot.1035

Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21(1):10–19; doi: 10.1093/bioinformatics/bth466

de la Cruz X, Hutchinson EG, Shepherd A, et al. Toward predicting protein topology: An approach to identifying beta hairpins. Proc Natl Acad Sci U S A2002;99(17):11157–11162; doi: 10.1073/pnas.162376199

Dhall A, Patiyal S, Raghava GPS. HLAncPred: A method for predicting promiscuous non-classical HLA binding sites. Brief Bioinform 2022;bbac192; In Press. doi: 10.1093/bib/bbac192

Dhall A, Patiyal S, Sharma N, et al. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. Brief Bioinform 2021;22(2):936–945; doi: 10.1093/bib/bbaa259

Dhanda SK, Singla D, Mondal AK, et al. DrugMint: A webserver for predicting and designing of drug-like molecules. Biol Direct 2013;8(1):28; doi: 10.1186/1745-6150-8-28

Dhanda SK, Usmani SS, Agrawal P, et al. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. Brief Bioinform 2017;18(3):467–478; doi: 10.1093/bib/bbw025

Dong J, Yao Z-J, Zhang L, et al. PyBioMed: A Python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Cheminform 2018;10(1):16; doi: 10.1186/s13321-018-0270-2

Emanuelsson O, Nielsen H, Brunak S, et al. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300(4):1005–1016; doi: 10.1006/jmbi.2000.3903

Faux N. Single amino acid and trinucleotide repeats: Function and evolution. Adv Exp Med Biol 2012;769:26–40; doi: 10.1007/978-1-4614-5434-2_3

Freeman TC, Wimley WC. TMBB-DB: A transmembrane β-barrel proteome database. Bioinformatics 2012;28(19): 2425–2430; doi: 10.1093/bioinformatics/bts478

Fuchs PFJ, Alix AJP. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 2005;59(4):828–839; doi: 10.1002/prot.20461

Garg A, Bhasin M, Raghava GPS. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 2005;280(15):14427–14432; doi: 10.1074/jbc.M411789200

Gupta S, Kapoor P, Chaudhary K, et al. In silico approach for predicting toxicity of peptides and proteins. PLoS One 2013;8(9):e73957; doi: 10.1371/journal.pone.0073957

Guralnik JM, LaCroix AZ, Branch LG, et al. Morbidity and disability in older persons in the years prior to death. Am J Public Health 1991;81(4):443–447; doi: 10.2105/ajph.81.4.443

Guruprasad K, Rajkumar S. Beta-and gamma-turns in proteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials. J Biosci 2000;25(2):143–156; doi: 10.1007/BF03404909

Hassan SS, Ghosh S, Attrish D, et al. Possible transmission flow of SARS-CoV-2 based on ACE2 features. Molecules 2020;25(24):5906; doi: 10.3390/molecules25245906

Horton P, Park K-J, Obayashi T, et al. WoLF PSORT: Protein localization predictor. Nucleic Acids Res 2007;35(Web Server): W585–W587; doi: 10.1093/nar/gkm259

Jahandideh S, Sarvestani AS, Abdolmaleki P, et al. Gamma-turn types prediction in proteins using the support vector machines. J Theor Biol 2007;249(4):785–790; doi: 10.1016/j.jtbi.2007.09.002

Jain S, Dhall A, Patiyal S, et al. IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. Comput Biol Med 2022;143:105297; doi: 10.1016/j.compbiomed.2022.105297

Kalita MK, Nandal UK, Pattnaik A, et al. CyclinPred: A SVM-based method for predicting cyclin protein sequences. PLoS One 2008;3(7):e2605; doi: 10.1371/journal.pone.0002605

Kalita MK, Ramasamy G, Duraisamy S, et al. ProtRepeatsDB: A database of amino acid repeats in genomes. BMC Bioinformatics 2006;7(1):336; doi: 10.1186/1471-2105-7-336

Kaundal R, Raghava GPS. RSLpred: An integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics 2009;9(9):2324–2342; doi: 10.1002/pmic.200700597

Kaur D, Arora C, Raghava GPS, et al. A hybrid model for predicting pattern recognition receptors using evolutionary information. Front Immunol 2020;11:71; doi: 10.3389/fimmu.2020.00071

Kaur H, Raghava GPS. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 2003a;12(5):923–929; doi: 10.1110/ps.0241703

Kaur H, Raghava GPS. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 2003b;12(3):627–634; doi: 10.1110/ps.0228903

Kaur H, Raghava GPS. A neural network method for prediction of β-turn types in proteins using evolutionary information. Bioinformatics 2004a;20(16):2751–2758; doi: 10.1093/bioinformatics/bth322

Kaur H, Raghava GPS. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins 2004b;55(1):83–90; doi: 10.1002/prot.10569

Kawashima S, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res 2000;28(1):374; doi: 10.1093/nar/28.1.374

Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res 1999;27(1):368–369; doi: 10.1093/nar/27.1.368

Kawashima S, Pokarowski P, Pokarowska M, et al. AAindex: Amino Acid Index Database, Progress Report 2008. Nucleic Acids Res 2008;36(Database issue):D202–D205; doi: 10.1093/nar/gkm998

Kountouris P, Hirst JD. Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures. BMC Bioinformatics 2010;11(1):407; doi: 10.1186/1471-2105-11-407

Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 2007;8(1); doi: 10.1186/1471-2105-8-463

Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 2008;71(1):189–194; doi: 10.1002/prot.21677

Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recogn 2011;24(2):303–313; doi: 10.1002/jmr.1061

Kumar R, Chaudhary K, Singh Chauhan J, et al. An in-Silico platform for predicting, screening and designing of antihypertensive peptides. Sci Rep 2015;5(1):12512; doi: 10.1038/srep12512

Kumar V, Agrawal P, Kumar R, et al. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. Front Microbiol 2018;9(APR); doi: 10.3389/fmicb.2018.00725

Kumar V, Kumar R, Agrawal P, et al. A method for predicting hemolytic potency of chemically modified peptides from its structure. Front Pharmacol 2020;11:54; doi: 10.3389/fphar.2020.00054

Kumar V, Patiyal S, Dhall A, et al. B3Pred: A random-forest-based method for predicting and designing blood–brain barrier penetrating peptides. Pharmaceutics 2021;13(8); doi: 10.3390/pharmaceutics13081237

Lata S, Mishra NK, Raghava GPS. AntiBP2: Improved version of antibacterial peptide prediction. BMC Bioinformatics 2010;11 Suppl 1(Suppl 1):S19; doi: 10.1186/1471-2105-11-S1-S19

Lata S, Raghava GPS. CytoPred: A server for prediction and classification of cytokines. Protein Eng Des Sel 2008;21(4):279–282; doi: 10.1093/protein/gzn006

Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides. BMC Bioinformatics 2007;8(1):263; doi: 10.1186/1471-2105-8-263

Li ZR, Lin HH, Han LY, et al. PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 2006;34(Web Server Issue):W32–W37; doi: 10.1093/nar/gkl305

Liu B. BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief Bioinform 2019;20(4):1280–1294; doi: 10.1093/bib/bbx165

Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res 2019;47(20):e127; doi: 10.1093/nar/gkz740

Luo H, Nijveen H. Understanding and identifying amino acid repeats. Brief Bioinform 2014;15(4):582–591; doi: 10.1093/bib/bbt003

Manavalan B, Basith S, Shin TH, et al. MLACP: Machine-learning-based prediction of anticancer peptides. Oncotarget 2017;8(44):77121–77136; doi: 10.18632/oncotarget.20365

Manavalan B, Basith S, Shin TH, et al. MAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2019;35(16):2757–2765; doi: 10.1093/bioinformatics/bty1047

Maryam L, Dhall A, Patiyal S, et al. Prediction of antibiotic resistant strains of bacteria from their beta-lactamases protein. biorxiv 2021; doi: 10.1101/2021.06.26.450028

Mathur D, Singh S, Mehta A, et al. In silico approaches for predicting the half-life of natural and modified peptides in blood. PLoS One 2018;13(6):e0196829; doi: 10.1371/journal.pone.0196829

McGinnis S, Madden TL. BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004;32(Web Server Issue):W20–W25; doi: 10.1093/nar/gkh435

Meher PK, Sahu TK, Saini V, et al. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's General PseAAC. Sci Rep 2017;7:1–12; doi: 10.1038/srep42362

Nagarajan R, Archana A, Thangakani AM, et al. PDBparam: Online resource for computing structural parameters of proteins. Bioinform Biol Insights 2016;10:73–80; doi: 10.4137/BBI.S38423

Nagel K, Jimeno-Yepes A, Rebholz-Schuhmann D. Annotation of protein residues based on a literature analysis: Cross-validation against UniProtKb. BMC Bioinformatics 2009;10 Suppl 8(S8):S4; doi: 10.1186/1471-2105-10-S8-S4

Nagpal G, Usmani SS, Dhanda SK, et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. Sci Rep 2017;7:42851; doi: 10.1038/srep42851

Ong SAK, Lin HH, Chen YZ, et al. Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics 2007;8(1):300; doi: 10.1186/1471-2105-8-300

Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinformatics 2013;14(1):44; doi: 10.1186/1471-2105-14-44

Patiyal S, Agrawal P, Kumar V, et al. NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. Protein Sci 2020;29(1):201–210; doi: 10.1002/pro.3761

Perutz MF, Johnson T, Suzuki M, et al. Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci U S A 1994;91(12):5355–5358; doi: 10.1073/pnas.91.12.5355

Rashid M, Saha S, Raghava GP. Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 2007;8(1):337; doi: 10.1186/1471-2105-8-337

Restrepo-Montoya D, Vizcaíno C, Niño LF, et al. Validating subcellular localization prediction tools with mycobacterial proteins. BMC Bioinformatics 2009;10(1):134; doi: 10.1186/1471-2105-10-134

Romero P, Obradovic Z, Li X, et al. Sequence complexity of disordered protein. Proteins 2001;42(1):38–48; doi: 10.1002/1097-0134(20010101)42:1 <38::aid-prot50>3.0.co;2-3

Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A 1993;90(16):7558–7562; doi: 10.1073/pnas.90.16.7558

Shahraki MF, Atanaki FF, Ariaeenejad S, et al. A computational learning paradigm to targeted discovery of biocatalysts from metagenomic data: A case study of lipase identification. Biotechnol Bioeng 2022;119(4):1115–1128; doi: 10.1002/bit.28037

Sharma A, Kapoor P, Gautam A, et al. Computational approach for designing tumor homing peptides. Sci Rep 2013;3(1):1607; doi: 10.1038/srep01607

Sharma N, Patiyal S, Dhall A, et al. AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. Brief Bioinform 2020;22(4):bbaa294. doi: 10.1093/bib/bbaa294

Singh H, Singh S, Raghava GPS. In silico platform for predicting and initiating β-turns in a protein at desired locations. Proteins 2015a;83(5):910–921; doi: 10.1002/prot.24783

Singh H, Singh S, Singla D, et al. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using random forest. Biol Direct 2015b;10(1):10; doi: 10.1186/s13062-015-0046-9

Singh S, Singh H, Tuknait A, et al. PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. Biol Direct 2015c;10(1):73; doi: 10.1186/s13062-015-0103-4

Singla D, Tewari R, Kumar A, et al. Designing of inhibitors against drug tolerant *Mycobacterium Tuberculosis* (H37Rv). Chem Centr J 2013;7(1):49; doi: 10.1186/1752-153X-7-49

Sofi MA, ArifWani M. Improving prediction of amyloid proteins using secondary structure-based alignments and segmented-PSSM. In: 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), Sematic Scholar, 2021; pp. 87–92.

Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res 2015;43(Database Issue): D364–D368; doi: 10.1093/nar/gku1028

Tyagi A, Kapoor P, Kumar R, et al. In silico models for designing and discovering novel anticancer peptides. Sci Rep 2013;3(1):2984; doi: 10.1038/srep02984

Usmani SS, Bedi G, Samuel JS, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One 2017;12(7); doi: 10.1371/journal.pone.0181748

Usmani SS, Bhalla S, Raghava GPS. Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. Front Pharmacol 2018a;9:954; doi: 10.3389/fphar.2018.00954

Usmani SS, Kumar R, Bhalla S, et al. In silico tools and databases for designing peptide-based vaccine and drugs. Adv Protein Chem Struct Biol 2018b;112:221–263; doi: 0.1016/bs.apcsb.2018.01.006

Verma R, Varshney GC, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids 2010;39(1):101–110; doi: 10.1007/s00726-009-0381-1

Wang Y, Xue Z, Xu J. Better prediction of the location of alpha-turns in proteins with support vector machine. Proteins 2006;65(1):49–54; doi: 10.1002/prot.21062

Wang J, Yang B, Revote J, et al. POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 2017;33(17):2756–2758; doi: 10.1093/bioinformatics/btx302

Xiao N, Cao D-S, Zhu M-F, et al. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31(11):1857–1859; doi: 10.1093/bioinformatics/btv042

Yap CW. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32(7):1466–1474; doi: 10.1002/jcc.21707

Address correspondence to:
Prof. Gajendra P.S. Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology
Okhla Phase 3
New Delhi 110020
India

E-mail: raghava@iiitd.ac.in