НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2024 Том 24 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2024 Vol. 24 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-1-101-111

Deep attention based Proto-oncogene prediction and Oncogene transition possibility detection using moments and position based amino acid features

Manickam Vijayalakshmi[1]✉, Mahesh Vallinayagi[2]

[1,2] Sri Sarada College for Women, Tirunelveli, 627011, India

[1] vijimarthresearch@gmail.com✉, https://orcid.org/0009-0001-2012-3169
[2] vallinayagimahesh@gmail.com, https://orcid.org/0009-0006-0552-0138

Abstract

The loss of the regulatory function of tumor suppression genes and mutations in Proto-oncogene are the common underlying mechanisms for uncontrolled tumor growth in the varied complex of disorders known as cancer. Oncogene can be curable by means of diagnosing and treating the possibilities of Proto-oncogene at earlier stages. Recently, machine learning approaches helps to focus and provide information about the possibilities of Proto-oncogene that may change into oncogene in different cancer types. This study helps to diagnose the possibilities of Proto-oncogene which are possible to change oncogenes at earlier stage. Thus, this present study proposed an efficient unique predictor of Proto-oncogene with the help of Bi-Directional Long Short Term Memory added with attention concept. This approach also find the probability of Proto-oncogene to oncogene using statistical moments, position based amino-acid composition representation and deep features extracted from the sequence. Consequently, this study suggests that using a K-Nearest Neighbor classifier it is possible to find probability of changing from Proto-oncogene to cancerous oncogene.

Keywords

Proto-oncogene, PseAAC, prediction, tumour suppression genes, TSG, machine learning, Bi-directional Long Short Term Memory (BiLSTM)

Acknowledgements

Special thanks to Dr. L. Rajagopala Marthandam, HOD of Medicine, TVMCH, India for his encouragement and support.

For citation: Vijayalakshmi M., Vallinayagi M. Deep attention based Proto-oncogene prediction and Oncogene transition possibility detection using moments and position based amino acid features. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 1, pp. 101–111. doi: 10.17586/2226-1494-2024-24-1-101-111

Основанное на особом интересе прогнозирование протоонкогена и обнаружение возможностей его мутации в онкоген на основе первоначального анализа последовательности аминокислот

Маникам Виджаялакшми[1]✉, Махеш Валлинаяги[2]

[1,2] Женский колледж Шри Сарада, Тирунелвели, 627011, Индия

[1] vijimarthresearch@gmail.com✉, https://orcid.org/0009-0001-2012-3169
[2] vallinayagimahesh@gmail.com, https://orcid.org/0009-0006-0552-0138

Аннотация

Утрата регуляторной функции генов, подавляющих опухоль, и мутации в протоонкогенах являются общими механизмами, лежащими в основе неконтролируемого роста опухолей при разнообразном комплексе заболеваний, известных как рак. Онкоген можно излечить путем диагностики и лечения возможностей протоонкогена на ранних стадиях. В последнее время подходы машинного обучения помогают сосредоточить внимание и предоставить информацию о возможностях протоонкогена, который может превращаться в онкоген при различных типах рака или изменять его на ранних стадиях. Предложен эффективный и уникальный предиктор протоонкогена с помощью нейронной сети Bi-Directional Long Short Term Memory (BiLSTM), дополненный концепцией ухода за больными. Этот подход также позволяет определить вероятность перехода от протоонкогена

© Vijayalakshmi M., Vallinayagi M., 2024

к онкогену с использованием статистических моментов, представления аминокислотного состава на основе положения и глубоких особенностей, извлеченных из последовательности. В работе применен классификатор K-Nearest Neighbor с помощью, которого можно определить вероятность перехода от протоонкогена к раковому онкогену.

Ключевые слова
протоонкогены, PseAAC, прогнозирование, гены опухолевой супрессии, TSG, машинное обучение, двунаправленная долговременная краткосрочная память, BiLSTM

Благодарности
Особая благодарность доктору Л. Раджагопале Мартандаму, руководителю медицины, TMCH, Индия, за его поощрение и поддержку.

Ссылка для цитирования: Виджаялакшми М., Валлинаяги М. Основанное на особом интересе прогнозирование протоонкогена и обнаружение возможностей его мутации в онкоген на основе первоначального анализа последовательности аминокислот // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 1. С. 101–111 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-1-101-111

Introduction

A series of nucleotide bases that make up each gene are responsible for carrying information about the development and operation of cells. This essentially happens when the cells transform the genetic code into proteins. In the human body, every protein has a particular purpose. Proto-oncogenes are common cellular genes that control cell division and development in humans [1]. It has been known for a long time that cancer is a result of loss of cell cycle control. The loss of control is a result of series of genetic mutations involving activation of Proto-oncogene to oncogenes and inactivation of tumour-suppressing genes.

The process of activation, which includes insertion mutations, point mutations, protein-protein interactions, retroviral transduction, gene amplification, chromosomal translocation, and transposon integration, can turn Proto-oncogene into oncogenes. Proto-oncogenes are frequently classed according to how closely their sequences resemble those of known proteins or according to how they typically behave inside cells [2]. Oncogenomics is the study of the genes linked to the development of cancer. By point mutations or gene amplification, Proto-oncogenes are frequently activated in transformed cells [3]. These genes may have a role in the genesis of cancer, and their identification may offer fresh perspectives on cancer treatment and diagnosis [4]. Oncogenes are thought to be distinguishable from other genes by identifying their unique mutation profile since the effects of mutations on genes activities are related to those effects [5]. Due to the significant heterogeneity of mutations across individuals and various cancer types, it is challenging to identify novel oncogenes aside from those that are often mutated [6]. Consequently, it is essential to create computational techniques for the finding.

Automatic protein functional observations have gained more interest lately because they narrow the search space for effective experimental annotation [7]. Various techniques, including prediction by sequence [8], protein-protein interactions [9], evolutionary relationships [10], protein structures and structure prediction algorithms [11], microarrays [12], and integration of data kinds [13], have enhanced tools for finding protein functional annotations. Additionally, a number of algorithms have been created to identify functional proteins from an amino acid sequence [14]. In general, protein functional detection research focuses on all types of functions, whether cancer-related or not. However, the subcategory of cancer-related functional detection is particularly helpful in cancer treatment. Hence, the present study has focused on which type of cancer is most possible along with protein functional detection for the given Proto-oncogene.

Related Reviews

The personalized therapy of cancer is a current research focus. This comprises a wide range of research projects centered on Proto-oncogene, oncogenes, DNA repair genes, DNA methylation, and tumour suppressor genes. For the purpose of identifying tumor suppressor genes in silico, several computational methods have been developed. Computer-based tools are capable of classifying complete protein functional detection as well as classifying various cancer types using clinical data, SNPs, and gene expressions in combination with conventional machine learning algorithms [15–20].

A given original protein sequence was utilized by Khan et al. [21] to extract location relative features for the identification of S-nitrosocystiene sites, which is the most common posttranslational modifications of proteins. In order to forecast Proto-oncogene, Malebary et al. [22] suggested statistical moments and position-based characteristics that were merged into Pseudo Amino-Acid Composition (PseAAC) based on Chou's 5-step rules and Random Forest (RF) classifier. A strategy for locating hydroxylysine sites was put out by Mahmood et al. [23] and it is based on a potent statistical and mathematical methodology that takes into account the sequence-order impact and the makeup of each item inside protein sequences. To ascertain if an amino acid substitution (AAS) affects protein function, the "Sorting Tolerant from Intolerant" (SIFT) method was employed in [24, 25]. Yang et al. [26] employed the word segmentation strategy to extract characteristics from the protein sequence. The characteristics were then classified using the Support Vector Machine (SVM).

Ali et al. [27] created a promising classification model with good membrane protein type discrimination. PseAAC is used to extract the silent characteristics of protein sequences. SVM, Nave Bayes, K-Nearest Neighbor, Voting Feature Interval, and Probabilistic Neural Network were used as classification techniques. A categorization system

for angiogenesis and cancer angiogenesis was put up by Allehaibi et al. [28]. Using a position- and composition-based method, variable-length proteome sequences were converted into fixed-length feature vectors. Utilizing statistical moments, position related information was further transformed into a condensed form. The best outcomes were determined using the three classifiers RF, Artificial Neural Network (ANN), and SVM. In order to detect TSGs and OGs by fusing extensive genetic and epigenetic data, Lyu et al. [29] created the algorithm Discovery of Oncogenes and Tumour SupressoR genes using Genetic and Epigenetic characteristics (DORGE). By incorporating nucleotide physicochemical characteristics into pseudo K-tuple nucleotide composition (PseKNC), Feng et al. [30] created a brand-new predictor known as iDNA6mA-PseKNC.

Huang et al. [31] established a technique to predict cancer proteins and used domain information to initially annotate protein interaction. Rahman et al. [32] introduced a system that directly extracts significant characteristics from protein sequences without relying on functional domain or structural information. They used the RF approach to rank the features after feature extraction and did the prediction scheme with SVM.

In order to determine a protein DNA-binding activity, Chowdhury et al. [33] created iDNAProt-ES, which makes use of both the evolutionary profile and structural data of proteins. They derived characteristics, such as amino acid composition, bigram, Dubchak features, auto-covariance, and segmentation distribution, from the Position-Specific Scoring Matrix (PSSM) profile. Ideal set of features are extracted using recursive feature elimination with the help of SPIDER2, the model was learned using SVM with a linear kernel.

Kumar et al. [34] exploit patient bias to find oncogenes far more effectively than current techniques by identifying it as a unique signal for cancer gene identification using RF classifier with relative/absolute position-based characteristics on Chou's PseAAC. Akmal et al. [35] suggested a unique predictor called iGlycoS-PseAAC.

Proto-oncogene to Oncogene Probability Score Detection Methodology ($PSD_{(p\text{-}o \rightarrow o)}$)

The proposed framework designed to find the probability of oncogene transformation from the given Proto-oncogene amino acid sequence; progress is clearly described in the following Fig. 1. This model provides relevant score for the chance of transforming Proto-oncogene into oncogene sequence in the type of breast, lung, kidney and collateral cancers or when the Proto-oncogene is stable as normal sequence. This approach extracts statistical moments and frequency and position based features [22] along with deep recurrent neural network of Bi-directional Long Short Term Memory (BiLSTM) features to find the chance of particular type of cancerous sequence formation or not with the help of traditional machine learning K-Nearest Neighbor classifier algorithm. This approach initially predicts whether a given sequence is Proto-oncogene or not with the help of BiLSTM network. Once it is identified as Proto-oncogene, it will check is there any possibility to be changed into oncogene.

Feature Extraction

This study utilized a variety of feature extraction strategies, such as Statistical Moments Calculation (raw, central, and Hahn), Position Relative Incidence Matrix (PRIM), Frequency Vector Determination, Absolute Position Incidence Vector (AAPIV) and Deep learning based feature with the help of BiLSTM.

Let peptide sample within the dataset be expressed as

$$\mathbf{PS}_l(Z) = S_0, S_1, \ldots, S_l,$$

where **PS** is Peptide Sample, Z contains the positive and negative samples, S_0, S_1… are the individual samples and l is non-uniform index indicating that the length of a sequence may vary. In other words, l represents the arbitrary length of the primary sequence which in this case is variable for each sample.

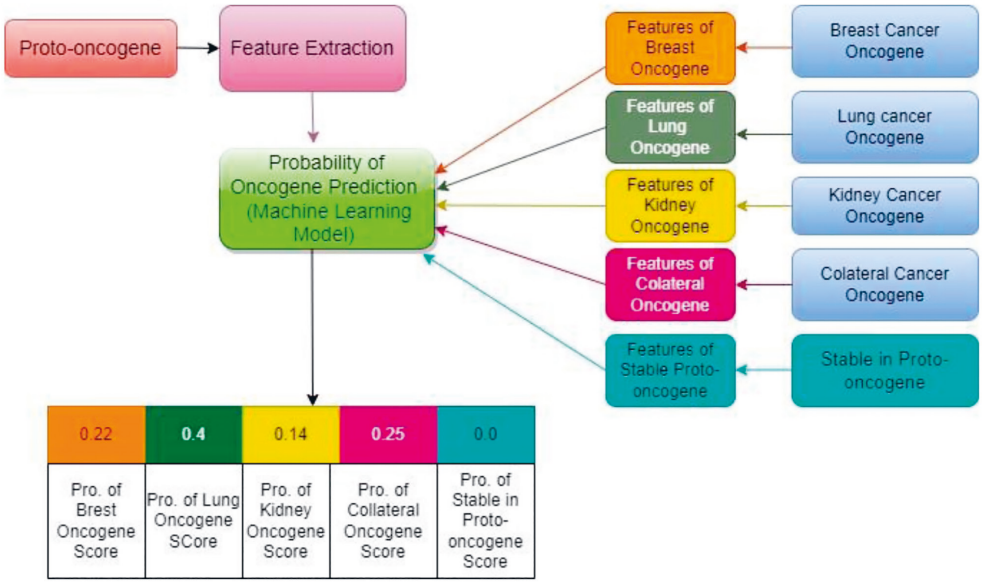

Fig. 1. The Architecture of $PSD_{(p\text{-}o \rightarrow o)}$ Model

Statistical Moments Calculation

The statistical moments are typically employed to extract particular qualities from data. Various moment sequences were employed to depict distinct attributes within the data. While some of these moments are useful for evaluating the direction and eccentricity of the data, others are useful for evaluating the magnitude of the data. There are several moment-defining polynomials depending on certain distributions have been proven by statisticians and mathematicians [36–40]. The suggested predictor raw moments, central moments, and Hahn moments were estimated up to order three. Ordering up to 3 generates enough information about the nature of data in numeric form [41].

Raw moments have location and scale variation features. Consequently, central moments are scale variant and location or position invariant. Orders up to three provide enough details on the type of data in numerical form. Additionally, the Hahn coefficient was determined using the Hahn polynomial which produces yet a different set of moments representing the initial data. The orthogonal features of these statistical moments led to their selection. The fact that orthogonal moments display a variety of features and may be utilized to recreate the original data means that they inherently include important properties that allow for exact categorization [42]. In general, these moments sufficiently transform information regarding the positioning and composition of residues in the primary structure.

The two-dimensional matrix **PS′** of size $n \times n$, which is a sequential transformation of all the amino acid residues of protein covered by **PS**, is the source of these moments computation (n is the order of the moment).

$$\mathbf{PS'} = \begin{bmatrix} ac_{11} & ac_{12} & \cdots & ac_{1n} \\ ac_{21} & ac_{22} & \cdots & ac_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ac_{n1} & ac_{n2} & \cdots & ac_{nn} \end{bmatrix}.$$

Using the function ω [43], **PS** is converted to **PS′**, and each of its arbitrary components, ac_{ij}, is an amino acid residue that has been placed in a two-dimensional setting. For all the moments up to degree 3, specific ordinal values of **PS′** elements were used. The raw moments are calculated by using equation:

$$MT_{ij} = \sum_{u=1}^{n} \sum_{v=1}^{n} u^i v^j ac_{uv},$$

where ac_{uv} is an arbitrary component of matrix **PS′**, and $i + j$ is the degree of the moments. Moreover, raw moments were denoted as MT_{00}, MT_{01}, MT_{02}, MT_{03}, MT_{10}, MT_{11}, MT_{12}, MT_{20}, MT_{21}, and MT_{30} for degree up to 3. The following equation is then used to determine central moments:

$$\eta_{ij} = \sum_{u=1}^{n} \sum_{v=1}^{n} (u - \bar{x})^i (v - \bar{y})^j ac_{uv},$$

where $\bar{x} = \dfrac{MT_{10}}{MT_{00}}$, and $\bar{y} = \dfrac{MT_{01}}{MT_{00}}$ which denotes the centroid of data.

PS was transformed into a square matrix **PS′** as it offers a substantial advantage for enumeration of Hahn moments. Discrete orthogonal moments in two dimensions require a square matrix as input. This orthogonal feature of Hahn moments suggests that they may be reversed using an inverse function. This reversible quality makes it easier to rebuild the data, which essentially means that it maintains the original data relative location, and sequence structure contained these moments.

For a one-dimensional matrix of size N, the Hahn polynomials of order n are calculated using the equation below.

$$h_n^{p,q}(r, N) = (N + V - 1)_n (N - 1)_n \times \sum_{k=0}^{n} (-1)^k \times$$

$$\times \frac{(-n)_k (-r)_k (2N + p + q - n - 1)_k}{(N + q - 1)_k (N - 1)_k} \frac{1}{k!},$$

where N is the size of the data array, V represents feature vector length n is the order of the moment; p and q are predefined constants. Additionally, the Pochhammer symbol q, which in turn employs the gamma operator as indicated in [41], is used in the equation. The two-dimensional Hahn moments are calculated using this Hahn coefficient as follows:

$$H_{ij} = \sum_{v=0}^{N-1} \sum_{u=0}^{N-1} ac_{uv} h_i^{p,q}(u, N).$$

The order of the moment is pointed out by the addition of i and j, that is, $i + j$; p, q are predefined constants; and ac_{uv} refers to any member in the square matrix **PS′**.

PRIM

To quantify the relative locations of amino acids and learn more about the relative positions of amino acid residues in the protein, a PRIM in the form of a 20×20 matrix was created. It is given as

$$\mathbf{PS_{PRIM}} = \begin{bmatrix} Seq_{1\to1} & Seq_{1\to2\ldots} & Seq_{1\to j\ldots} & Seq_{1\to20} \\ Seq_{2\to1} & Seq_{2\to2\ldots} & Seq_{2\to j\ldots} & Seq_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Seq_{i\to1} & Seq_{i\to2\ldots} & Seq_{i\to j\ldots} & Seq_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Seq_{20\to1} & Seq_{20\to2\ldots} & Seq_{20\to j\ldots} & Seq_{20\to20} \end{bmatrix}.$$

The sum of the positions of the j^{th} theresidue and the i^{th} residue initial occurrence for each element of this matrix is represented by the symbol $Seq_{i\to j}$. As a result, this matrix has 400 coefficients, which is a very large number. The opportunity to condense this information into a concise form is made possible by statistical moments. There are 30 coefficients total for degrees up to 3 after computing the PRIM raw, central, and Hahn moments. In the same way, the Reverse Position Relative Incidence Matrix (RPRIM) was created using a basic protein sequence.

The PRIM can be denoted as:

$$\mathbf{PS_{RPRIM}} = \begin{bmatrix} Seq_{1\to1} & Seq_{1\to2\ldots} & Seq_{1\to j\ldots} & Seq_{1\to20} \\ Seq_{2\to1} & Seq_{2\to2\ldots} & Seq_{2\to j\ldots} & Seq_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Seq_{i\to1} & Seq_{i\to2\ldots} & Seq_{i\to j\ldots} & Seq_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Seq_{20\to1} & Seq_{20\to2\ldots} & Seq_{20\to j\ldots} & Seq_{20\to20} \end{bmatrix}. \quad (1)$$

Statistical moments were used to decrease the dimensionality of RPRIM, resulting in the construction of a set of 30 elements. The PRIM describes the relative positions of amino acid residues in a polypeptide chain. This information is augmented by the RPRIM in (1), which reveals even more hidden information by repeating the operation on the opposite side of the primary sequence.

Frequency Vector Determination

Simple counting of each amino acid residue inside the main sequence yields the frequency vector. The frequency of each amino acid residue in the supplied sequence is represented by an element in the frequency vector. As a result, the frequency vector has 20 coefficients.

$$\mathbf{fv} = \{\propto_1, \propto_2 \ldots \ldots \propto_{20}\}.$$

Absolute Position Incidence Vector

The frequency matrix significantly provides information on how amino acid residues are composed. The AAPIV gives a summary of the residue location. It had a length of 20 elements and was composed of a single coefficient for each amino acid residue. The sum of the positions of every natural amino acid in the fundamental structure is included in elements of the AAPIV, which is given as equation:

$$\mathbf{PO} = \{pos_1, pos_2, pos_3, \ldots, pos_{20}\}.$$

Equation below made it possible to calculate any i^{th} element of the AAPIV.

$$pos_i = \sum_{PO=1}^{n} pos_k,$$

where pos_k represents the location of the i^{th} amino acid residue. Consequently, the Reverse Accumulative Absolute Position Vector (RAAPIV) assessed additional specific data based on the absolute positions of amino acids in peptide samples. RAAPIV was produced by reversing the basic sequence and calculating AAPIV. That can be denoted as

$$\mathbf{RAAPIV} = \{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{20}\},$$

where λ_i is the total number of positions in the main structure; then the i^{th} amino acid residue can be found.

BiLSTM

The architecture of BiLSTM is depicted in Fig. 2. The input sequences will be routed into the BiLSTM with size 128 filter. The outcome of BiLSTM is sent into maxpooling layer MP1 with size 2. Then, the Relu Layer will be applied to the output. The output of the Relu layer is then fed into the attention layer. The dropout layer will be given the results of the attention layer. The output of the BiLSTM layers is fed into the maxpooling layer MP2 with size 2. The outcome will then be applied to the Relu layer and then into the dense layer with size 128. The SoftMax classification will be used in the dense output to predict whether a given sequence is Proto-oncogene or not at the end of the architecture shown in the Fig. 2. The same layer model is used for extracting deep features as well as for the prediction of Proto-oncogene.

Feature Vector Description

The final step in processing primary sequences (**PS′**) through all of the aforementioned phases is combining them to create an accumulative feature vector. Two dimensional representation of the major sequence matrices **PRIM**, **PS′**, and **RPRIM** are changed into a concise form through calculating their statistical moments (raw, central, and Hahn). Thus, it produces 90 coefficients. Another 60 coefficients are included to the vector by pooling the frequency vector (**fv**), AAPIV (**PO**), and RAAPIV (**RAAPIV**). The deep features extracted by the final dense layer of the model with size of 128. Hence the final feature vector of the size 278 is used to find whether the given Proto-oncogene will change into oncogene or not.

Probability Score Detection using ML Algorithm

After features were extracted using the feature extraction approaches, a fixed-sized feature vector with 278 coefficients was created to be used for further processing in the machine learning algorithm for computing the possibilities of oncogene prediction. The features of Proto-oncogene sequences were extracted by utilizing certain

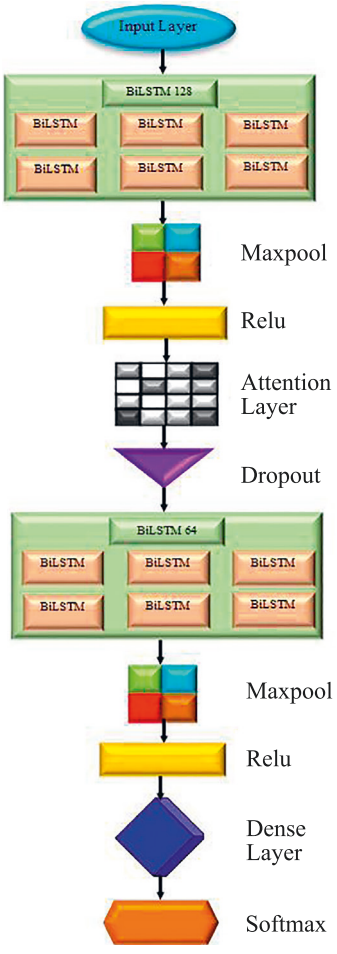

Fig. 2. BiLSTM Architecture

feature extraction strategies like Statistical Moments Calculation (raw, central, and Hahn), PRIM, Frequency Vector Determination, AAPIV, and they are fed into the K-Nearest Neighbor classifier Machine learning model which predicts the probability of the formation of Oncogene that causes cancer. Furthermore, the different categories of oncogene sequences, like breast cancer oncogene, lung cancer oncogene, kidney cancer oncogene, collateral oncogene and stable Proto-oncogene, were extracted by the same feature extraction strategies. All the above feature vectors are fed into the proposed oncogene probability prediction machine learning model. This proposed model compares the available features to identify the probability score of the Proto-oncogene to mutate into different types of oncogene. It also predicts the probability score of Proto-oncogene sample tested is a stable Proto-oncogene, which means it does not mutates into an oncogene.

Experimental Results

Dataset Description

The UniProt website consists of different kind resources, such as UniProt Knowledge Base (UniProtKB), UniProt Reference (UniRef), UniProt Archives (UniParc), and Protein Sets (Proteomes) of Fully Sequenced Genomes. Supporting datasets include protein information, such as transcript, distribution, sub cellular location, keywords, cross reference datasets, and disease, currently available in the UniProtKB protein entry. With the help of UniProt's "Search/ID Mapping" tool, different described in UniProt [43]. The following Fig. 3 shows the UniProt search and downloading section for genomic data. In this work, the 630 negative samples and 252 positive samples from the dataset provided by the ProtoPred [22] are used to train the BiLSTM with attention model for Proto-oncogene prediction as well as those 252 positive samples are used to find the probability of its status from Proto-oncogene to oncogene which causes different types of cancer. The following Table 1 shows the sample Proto-oncogene [22] and oncogenes in different types of cancer from UniProtKB [43].

The benchmark data set is split into k (10) disjoint fold partitions for cross-validation. Table 2 displays the findings of the KFold cross validations of the proposed model and the following Table 3 shows the performance of the Proto-oncogene prediction using the designed BiLSTM with attention model compared with state art of works, as an independent test, 30 % samples used for testing and remaining 70 % of samples used for training.

From the Table 3 it is found that the proposed BiLSTM_ATT model archives 97 % F1-Score, which is significantly better compared to the existing approaches. In the second phase of Proto-oncogene to oncogene transformation probability finding process, we trained the KNN classifier model with all the oncogene from different types of disease features along with stable Proto-oncogene features.

In the evaluation progress, for all of the 252 Proto-oncogene the same set of statistical moments, frequency based features along with BiLSTM features are extracted and score for each five classes, such as breast cancer, lung cancer, kidney cancer, colorectal cancer and the Proto-oncogene remains in the same state as denoted by stable, is estimated. Among the five scores, the class belonging to the maximum score is considered as the probability of changing Proto-oncogene to that particular type of oncogene or shows that there is no transition. The following Pie chart in the Fig. 4 clearly describes that the percentage of Proto-oncogene in the dataset has the probability of changing into particular types of cancer disease or it won't affect anything. A Pie chart is drawn based on the probability score attained by each Proto-oncogene sequence in the benchmark dataset [22], and the highest score identifies the type of oncogene that is most probable.

From the execution of the design, it is found that the up to 43.3 % of sequence data has the probability of changing into breast cancer oncogene. Similarly, 30 % related to kidney cancer, 13.3 % possibility of colorectal cancer, 10.8 % of lung cancer and 3.6 % of no transition. In this mode the probability of changes is estimated based on the highest score among all 5 classes. The highest score may be in any range from zero to one. In order to estimate the

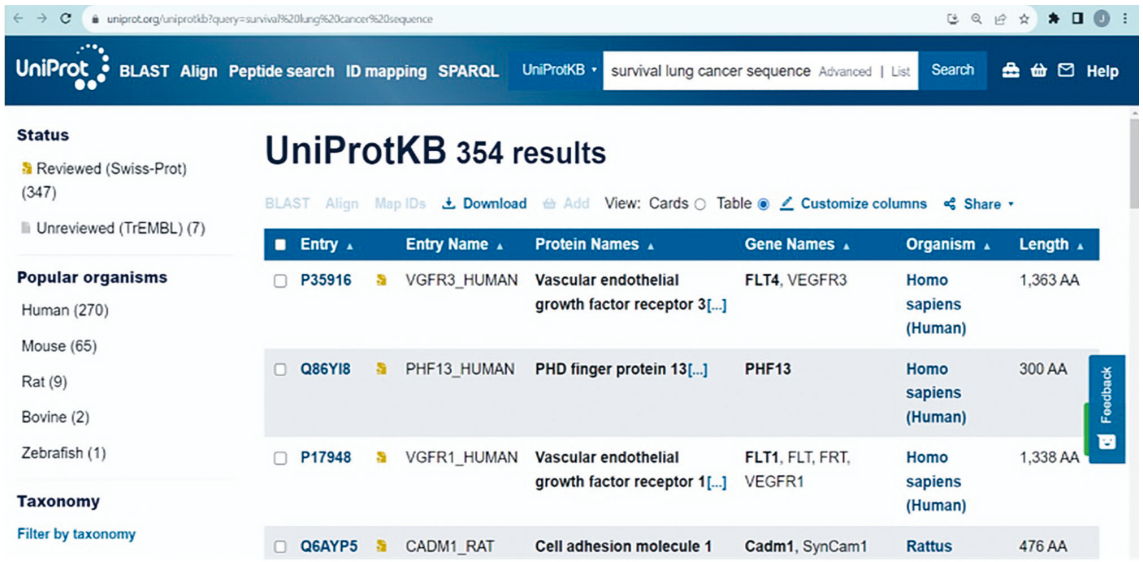

Fig. 3. UniProtKBSearch and its Result Page

Table 1. Samples studied

Sample of Proto-oncogene sequence [22]
MECPSCQHVSKEETPKFCSQCGERLPPAAPIADSENNNSTMASA…
MAHSCRWRFPARPGTTGGGGGGGRRGLGGAPRQRVPALLLPPGP…
MKLNPQQAPLYGDCVVTVLLAEEDKAEDDVVFYLVFLGSTLRHC…
MDQTCELPRRNCLLPFSNPVNLDAPEDKDSPFGNGQSNFSEPLN…
MTSGGSASRSGHRGVPMTSRGFDGSRRGSLRRAGARETASEAAD…
Sample of the Kidney Cancer Sequence
MFASCHCVPRGRRTMKMIHFRSSSVKSLSQEMRCTIRLLDDSEISCHI…
MGQDAFMEPFGDTLGVFQCKIYLLLFGACSGLKVTVPSHTVHGVRG…
MRLEELKRLQNPLEQVNDGKYSFENHQLAMDAENNIEKYPLNLQPL…
MTVKTEAAKGTLTYSRMRGMVAILIAFMKQRRMGLNDFIQKIANN…
MAEQDVENDLLDYDEEEEPQAPQESTPAPPKKDIKGSYVSIHSSGFR…
Sample of the Lung Cancer Sequence
MENEKENLFCEPHKRGLMKTPLKESTTANIVLAEIQPDFGPLTTP…
MENFTALFGAQADPPPPPTALGFGPGKPPPPPPPAGGGPGTA…
MKIIILLGFLGATLSAPLIPQRLMSASNSNELLLNLNNGQLLPLQL…
MPVSTSLHQDGSQERPVSLTSTTSSSGSSCDSRSAMEEPSSSEA…
MAFSDLTSRTVHLYDNWIKDADPRVEDWLLMSSPLPQTILLGF…
Sample of the Breast Cancer Sequence
MGQDAFMEPFGDTLGVFQCKIYLLLFGACSGLKVTVPSHTVHG…
MVQYELWAALPGASGVALACCFVAAAVALRWSGRRTARGAV…
MNYSLHLAFVCLSLFTERMCIQGSQFNVEVGRSDKLSLPGFENL…
MAGFGAMEKFLVEYKSAVEKKLAEYKCNTNTAIELKLVRFPEDL…
MDRSKENCISGPVKATAPVGGPKRVLVTQQFPCQNPLPVNSG…
Sample of the Colorectal Cancer Sequence
MKIIILLGFLGATLSAPLIPQRLMSASNSNELLLNLNNGQLLPLQ…
MSEKPKVYQGVRVKITVKELLQQRRAHQAASGGTRSGGSSVH…
MELSGATMARGLAVLLVLFLHIKNLPAQAADTCPEVKVVGLEG…
MIPPADSLLKYDTPVLVSRNTEKRSPKARLLKVSPQQPGPSGSA…
MEGAALLRVSVLCIWMSALFLGVGVRAEEAGARVQQNVPSGT…

Table 2. Findings of the KFold testing for Accuracy and F1-Score, %

Fold #	Accuracy	Precision	Recall	F1-Score
1	94	96	90	92
2	97	98	96	97
3	94	96	90	92
4	96	96	95	95
5	94	96	90	92
6	97	97	97	97
7	95	97	92	94
8	97	98	96	97
9	98	99	98	98
10	100	100	100	100
Average	96	96	94	95

chance of changing from Proto-oncogene to oncogene in reliable form, a marginal threshold is fixed for the maximum score checking. In this work the threshold is fixed as 0.5. Here the chance of oncogene cancer type is determined not only by its maximum score, it also checks if that maximum score is greater than the pre-defined threshold. The probability of changes from Proto-oncogene to particular four types of cancerous gene or stable in the current state is shown in the following Fig. 5 with the threshold as 0.5. The predicted data whose highest score below 0.5 means it is in unpredictable class. Whereas in Fig. 4, the data is assigned to the class which has highest predicted score among all trained five classes.

From the above Fig. 5, it is clear that the based on the threshold the probability of changing from Proto-oncogene to breast cancer oncogene is reduced from 43.3 % without threshold (Fig. 4) to 35.4 %. The following Fig. 6 shows the probability score for changing from Proto-oncogene to oncogene for the given testing unknown Proto-oncogene sequence in the Table 4.

Table 3. Results of Proto-oncogene prediction using BiLSTM with Spatial Attention, %

Methods	Accuracy	F1-Score	Precision	Recall	Specificity	Mcc	AUC
PSSM [44]	81	78	77	79	79	56	83
PseAAC [45]	85	82	81	83	83	64	89
ProtoPred_RF [22]	97	96	94	98	98	92	97
BiLSTM_ATT Model	97	97	96	98	98	94	98

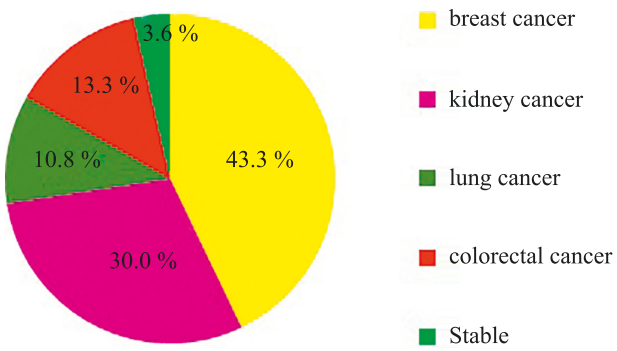

Fig. 4. Percentage of Proto-oncogene has the possibility to change into particular type of oncogene or remains as it as Proto-oncogene

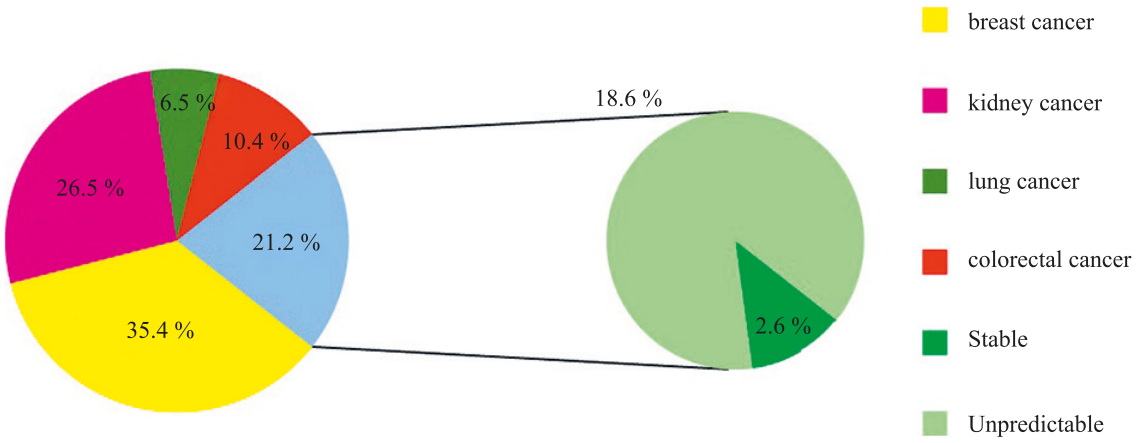

Fig. 5. Probability of Transition from Proto-oncogene to Oncogenes with Score Threshold

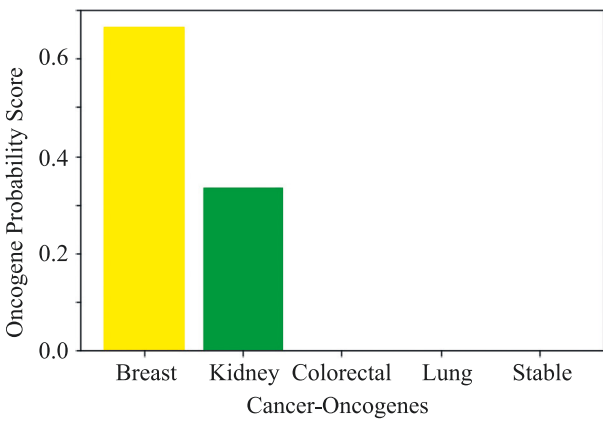

Fig. 6. Predicted Probability Score for the Cancer Oncogene

Table 4. Unknown Proto-oncogene Sequence for Prediction

MEYMSTGSDEKEEIDLLIKHLNVSEVIDIMENLYASEEPGVYEPSLMTMYPDS
NQNEERSESLLRSGQEVPWLSSVRYGTVEDLLAFANHVSNMTKHFYGRRPQ
ECGILLNMVISPQNGRYQIDSDVLLVPWKLTYRNIGSGFVPRGAFGKVYLA
QDMKTKKRMACKLIPIDQFKPSDVEIQACFRHENIAELYGAVLWGDTVHLFM
EAGEGGSVLEKLESCGPMREFEIIWVTKHILKGLDFLHSKKVIHHDIKPSNIV
FMSTKAVLVDFGLSVKMTEDVYLPKDLRGTEIYMSPEVILCRGHSTKADIY
SLGATLIHMQTGTPPWVKRYPRSAYPSYLYIIHKQAPPLEDIAGDCSPGMRELI
EAALERNPNHRPKAADLLKHEALNPPREDQPRCQSLDSALFERKRLLSRKEL
QLPENIADSSCTGSTEESEVLRRQRSLYIDLGALAGYFNIVRGPPTLEYG

Conclusion

Mutations in Proto-oncogene are the leading causes of cancer because of exposure to a mutagen. Proto-oncogene proteins are formed when Proto-oncogene are translated. These proteins function as a biomarker for cancer susceptibility. The proposed approach offers a reliable in-silico method for detecting such proteins. The suggested method incorporates all of the suggestions from the state of the art to create a computationally intelligent predictor. The features of a two-dimensional representation of the key structure of proteins, such as Statistical Moment Calculation (raw, central, and Hahn), Position Relative Incidence Matrix, Frequency Vector Determination, Absolute Position Incidence Vector and Deep RNN features, are gathered to form feature vectors. Following the extraction of feature vectors from both positive and negative sequences, the data is used to train the K-Nearest Neighbor machine learning classifier algorithm that is employed to find the probability score for each classes, such as stable, unpredictable or breast, lung, kidney and collateral cancers. The obtained results are evaluated using the benchmark ProroPred dataset. They show that the designed BiLSTM_Attention model achieves 97 % accuracy for the prediction of Proto-oncogene. The deep feature extraction along with statistical and moments features support the model to find the probability of transformation from Proto-oncogene to oncogene. This approach helps to find transformation from of Proto-oncogene to oncogene at earlier stage, which saves human life.

References

1. Williams D.E., Eisenman J., Baird A., Rauch C., Van Ness K., March C.J., Park L.S., Martin U., Mochizukl D.Y., Boswell H.S., Burgess G.S., Cosman D., Lyman S.D. Identification of a ligand for the c-kit Proto-oncogene. *Cell*, 1990, vol. 63, no. 1, pp. 167–174. https://doi.org/10.1016/0092-8674(90)90297-r
2. Cooper G.M. *Oncogenes*. 2nd ed. Jones and Bartlett Publishers Inc. Boston, 1995, 384 p.
3. Mulligan L.M., Kwok J.B., Healey C.S., Elsdon M.J., Eng C., Gardner E., Love D.R., Mole S.E., Moore J.K., Papi L., Ponder M.A., Telenius H., Tunnacliffe A., Ponder B.A. Germ-line mutations of the *RET* Proto-oncogene in multiple endocrine neoplasia type 2A. *Nature,* 1993, vol. 363, no. 6428, pp. 458–460. https://doi.org/10.1038/363458a0
4. Croce C.M. Oncogenes and cancer. *New England journal of medicine*, 2008, vol. 358, no. 5, pp. 502–511. https://doi.org/10.1056/NEJMra072367
5. Vogelstein B., Papadopoulos N., Velculescu V.E., Diaz L.A., Kinzler K.W. Cancer genome landscapes. *Science*, 2013, vol. 339, no. 6127, pp. 1546–1558. https://doi.org/10.1126/science.1235122
6. Pon J.R., Marra M.A. Driver and passenger mutations in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 2015, vol. 10, pp. 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312
7. Kulmanov M., Khan M.A., Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 2018, vol. 34, no. 4, pp. 660–668. https://doi.org/10.1093/bioinformatics/btx624
8. Wass M.N., Sternberg M.J. ConFunc–functional annotation in the twilight zone. *Bioinformatics*, 2008, vol. 24, no. 6, pp. 798–806. https://doi.org/10.1093/bioinformatics/btn037
9. Deng M., Zhang K., Mehta S., Chen T., Sun F. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 2003, vol. 10, no. 6, pp. 947–960. https://doi.org/10.1089/106652703322756168
10. Marcotte E.M., Pellegrini M., Ng H.L., Rice D.W., Yeates T.O., Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 1999, vol. 285, no. 5428, pp. 751–753. https://doi.org/10.1126/science.285.5428.751

Литература

1. Williams D.E., Eisenman J., Baird A., Rauch C., Van Ness K., March C.J., Park L.S., Martin U., Mochizukl D.Y., Boswell H.S., Burgess G.S., Cosman D., Lyman S.D. Identification of a ligand for the c-kit Proto-oncogene // Cell. 1990. V. 63. N 1. P. 167–174. https://doi.org/10.1016/0092-8674(90)90297-r
2. Cooper G.M. Oncogenes / 2nd ed. Jones and Bartlett Publishers Inc. Boston, 1995. 384 p.
3. Mulligan L.M., Kwok J.B., Healey C.S., Elsdon M.J., Eng C., Gardner E., Love D.R., Mole S.E., Moore J.K., Papi L., Ponder M.A., Telenius H., Tunnacliffe A., Ponder B.A. Germ-line mutations of the RET Proto-oncogene in multiple endocrine neoplasia type 2A // Nature. 1993. V. 363. N 6428. P. 458–460. https://doi.org/10.1038/363458a0
4. Croce C.M. Oncogenes and cancer // New England journal of medicine. 2008. V. 358. N 5. P. 502–511. https://doi.org/10.1056/NEJMra072367
5. Vogelstein B., Papadopoulos N., Velculescu V.E., Diaz L.A., Kinzler K.W. Cancer genome landscapes // Science. 2013. V. 339. N 6127. P. 1546–1558. https://doi.org/10.1126/science.1235122
6. Pon J.R., Marra M.A. Driver and passenger mutations in cancer // Annual Review of Pathology: Mechanisms of Disease. 2015. V. 10. P. 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312
7. Kulmanov M., Khan M.A., Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier // Bioinformatics. 2018. V. 34. N 4. P. 660–668. https://doi.org/10.1093/bioinformatics/btx624
8. Wass M.N., Sternberg M.J. ConFunc–functional annotation in the twilight zone // Bioinformatics. 2008. V. 24. N 6. P. 798–806. https://doi.org/10.1093/bioinformatics/btn037
9. Deng M., Zhang K., Mehta S., Chen T., Sun F. Prediction of protein function using protein-protein interaction data // Journal of Computational Biology. 2003. V. 10. N 6. P. 947–960. https://doi.org/10.1089/106652703322756168
10. Marcotte E.M., Pellegrini M., Ng H.L., Rice D.W., Yeates T.O., Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences // Science. 1999. V. 285. N 5428. P. 751–753. https://doi.org/10.1126/science.285.5428.751

11. Pal D., Eisenberg D. Inference of protein function from protein structure. *Structure*, 2005, vol. 13, no. 1, pp. 121–130. https://doi.org/10.1016/j.str.2004.10.015
12. Huttenhower C., Hibbs M., Myers C., Troyanskaya O.G. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 2006, vol. 22, no. 23, pp. 2890–2897. https://doi.org/10.1093/bioinformatics/btl492
13. Kourmpetis Y.A.I., van Dijk A.D.J., Bink M.C.A., van Ham M.R.C.H.J., terBraak C.J.F. Bayesian markov random field analysis for protein function prediction based on network data. *PLoS One*, 2010, vol. 5, no. 2. https://doi.org/10.1371/journal.pone.0009293
14. Radivojac P., Clark W.T., Oron T.R. et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 2013, vol. 10, no. 3, pp. 221–227. https://doi.org/10.1038/nmeth.2340
15. Mihaylov I., Nisheva M., Vassilev D. Application of machine learning models for survival prognosis in breast cancer studies. *Information*, 2019, vol. 10, no. 3, pp. 93. https://doi.org/10.3390/info10030093
16. Cruz J.A., Wishart D.S. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2006, vol. 2, pp. 59–77. https://doi.org/10.1177/117693510600200030
17. Sotiriou C., Neo S.-Y., McShane L.M., Korn E.L., Long P.M., Jazaeri A., Martiat P., Fox S.B., Harris A.L., Liu E.T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, vol. 100, no. 18, pp. 10393–10398. https://doi.org/10.1073/pnas.1732912100
18. Vural S., Wang X., Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Systems Biology*, 2016, vol. 10, no. 3, pp. 62. https://doi.org/10.1186/s12918-016-0306-z
19. Cai Z., Xu D., Zhang Q., Zhang J., Ngai S.-M., Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 2015, vol. 11, no. 3, pp. 791–800. https://doi.org/10.1039/c4mb00659c
20. Kourou K., Exarchos T.P., Exarchos K.P., Karamouzis M.V. Fotiadis D.I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 2015, vol. 13, pp. 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
21. Khan Y.D., Batool A., Rasool N., Khan S.A., Chou K.-C.J. Prediction of nitrosocysteine sites using position and composition variant features. *Letters in Organic Chemistry*, 2019, vol. 16, no. 4, pp. 283–293. https://doi.org/10.2174/1570178615666180802122953
22. Malebary S.J., Khan R., Khan Y.D. ProtoPred: Advancing oncological research through identification of proto-oncogene proteins. *IEEE Access*, 2021, vol. 9, pp. 68788–68797. https://doi.org/10.1109/ACCESS.2021.3076448
23. Mahmood M.K., Ehsan A., Khan Y.D., Chou K.-C. iHyd-LysSite (EPSV): identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique. *Current Genomic*, 2020, vol. 21, no. 7, pp. 536–545. https://doi.org/10.2174/1389202921999200831142629
24. Kumar P., Henikoff S., Ng P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 2009, vol. 4, no. 7, pp. 1073–1081. https://doi.org/10.1038/nprot.2009.86
25. Vaser R., Adusumalli S., Leng S., Sikic M., Ng P.C. SIFT missense predictions for genomes. *Nature Protocols*, 2016, vol. 11, no. 1, pp. 1–9. https://doi.org/10.1038/nprot.2015.123
26. Yang Y., Lu B.L., Yang W.Y. Classification of protein sequences based on word segmentation methods. *Proc. of the 6th Asia-Pacific Bioinformatics Conference (APBC '08)*, 2008, pp. 177–186. https://doi.org/10.1142/9781848161092_0020
27. Ali F., Hayat M. Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, 2015, vol. 384, pp. 78–83. https://doi.org/10.1016/j.jtbi.2015.07.034
28. Allehaibi K., Daanial Khan Y., Khan S.A. iTAGPred: A two-level prediction model for identification of angiogenesis and tumor angiogenesis biomarkers. *Applied Bionics and Biomechanics*, 2021, vol. 2021, pp. 2803147. https://doi.org/10.1155/2021/2803147
29. Lyu J., Li J.J., Su J., Peng F., Chen Y.E., Ge X., Li W. DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. *Science Advances*, 2020, vol. 6, no. 46, pp. 1–17. https://doi.org/10.1126/sciadv.aba6784
30. Feng P., Yang H., Ding H., Lin H., Chen W., Chou K.C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by

11. Pal D., Eisenberg D. Inference of protein function from protein structure // Structure. 2005. V. 13. N 1. P. 121–130. https://doi.org/10.1016/j.str.2004.10.015
12. Huttenhower C., Hibbs M., Myers C., Troyanskaya O.G. A scalable method for integration and functional analysis of multiple microarray datasets // Bioinformatics. 2006. V. 22. N 23. P. 2890–2897. https://doi.org/10.1093/bioinformatics/btl492
13. Kourmpetis Y.A.I., van Dijk A.D.J., Bink M.C.A., van Ham M.R.C.H.J., terBraak C.J.F. Bayesian markov random field analysis for protein function prediction based on network data // PLoS One. 2010. V. 5. N 2. https://doi.org/10.1371/journal.pone.0009293
14. Radivojac P., Clark W.T., Oron T.R. et al. A large-scale evaluation of computational protein function prediction // Nature Methods. 2013. V. 10. N 3. P. 221–227. https://doi.org/10.1038/nmeth.2340
15. Mihaylov I., Nisheva M., Vassilev D. Application of machine learning models for survival prognosis in breast cancer studies // Information. 2019. V. 10. N 3. P. 93. https://doi.org/10.3390/info10030093
16. Cruz J.A., Wishart D.S. Applications of machine learning in cancer prediction and prognosis // Cancer Informatics. 2006. V. 2. P. 59–77. https://doi.org/10.1177/117693510600200030
17. Sotiriou C., Neo S.-Y., McShane L.M., Korn E.L., Long P.M., Jazaeri A., Martiat P., Fox S.B., Harris A.L., Liu E.T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study // Proceedings of the National Academy of Sciences of the United States of America. 2003. V. 100. N 18. P. 10393–10398. https://doi.org/10.1073/pnas.1732912100
18. Vural S., Wang X., Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches // BMC Systems Biology. 2016. V. 10. N 3. P. 62. https://doi.org/10.1186/s12918-016-0306-z
19. Cai Z., Xu D., Zhang Q., Zhang J., Ngai S.-M., Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods // Molecular BioSystems. 2015. V. 11. N 3. P. 791–800. https://doi.org/10.1039/c4mb00659c
20. Kourou K., Exarchos T.P., Exarchos K.P., Karamouzis M.V. Fotiadis D.I. Machine learning applications in cancer prognosis and prediction // Computational and Structural Biotechnology Journal. 2015. V. 13. P. 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
21. Khan Y.D., Batool A., Rasool N., Khan S.A., Chou K.-C.J. Prediction of nitrosocysteine sites using position and composition variant features // Letters in Organic Chemistry. 2019. V. 16. N 4. P. 283–293. https://doi.org/10.2174/1570178615666180802122953
22. Malebary S.J., Khan R., Khan Y.D. ProtoPred: Advancing oncological research through identification of proto-oncogene proteins // IEEE Access. 2021. V. 9. P. 68788–68797. https://doi.org/10.1109/ACCESS.2021.3076448
23. Mahmood M.K., Ehsan A., Khan Y.D., Chou K.-C. iHyd-LysSite (EPSV): identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique // Current Genomic. 2020. V. 21. N 7. P. 536–545. https://doi.org/10.2174/1389202921999200831142629
24. Kumar P., Henikoff S., Ng P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm // Nature Protocols. 2009. V. 4. N 7. P. 1073–1081. https://doi.org/10.1038/nprot.2009.86
25. Vaser R., Adusumalli S., Leng S., Sikic M., Ng P.C. SIFT missense predictions for genomes // Nature Protocols. 2016. V. 11. N 1. P. 1–9. https://doi.org/10.1038/nprot.2015.123
26. Yang Y., Lu B.L., Yang W.Y. Classification of protein sequences based on word segmentation methods // Proc. of the 6th Asia-Pacific Bioinformatics Conference (APBC '08). 2008. P. 177–186. https://doi.org/10.1142/9781848161092_0020
27. Ali F., Hayat M. Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition // Journal of Theoretical Biology. 2015. V. 384. P. 78–83. https://doi.org/10.1016/j.jtbi.2015.07.034
28. Allehaibi K., Daanial Khan Y., Khan S.A. iTAGPred: A two-level prediction model for identification of angiogenesis and tumor angiogenesis biomarkers // Applied Bionics and Biomechanics. 2021. V. 2021.P. 2803147. https://doi.org/10.1155/2021/2803147
29. Lyu J., Li J.J., Su J., Peng F., Chen Y.E., Ge X., Li W. DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features // Science Advances. 2020. V. 6. N 46. P. 1–17. https://doi.org/10.1126/sciadv.aba6784
30. Feng P., Yang H., Ding H., Lin H., Chen W., Chou K.C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by

incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, 2018, vol. 111, no. 1, pp. 96–102. https://doi.org/10.1016/j.ygeno.2018.01.005
31. Huang C.H., Peng H.S., Ng K.L. Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms. *BioMed Research International*, 2015, vol. 2015, pp. 312047. https://doi.org/10.1155/2015/312047
32. Rahman M.S., Shatabda S., Saha S., Kaykobad M., Rahman M.S. DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC. *Journal of Theoretical Biology*, 2018, vol. 452, pp. 22–34. https://doi.org/10.1016/j.jtbi.2018.05.006
33. Chowdhury S.Y., Shatabda S., Dehzangi A. iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features. *Scientific Reports*, 2017, vol. 7, pp. 14938. https://doi.org/10.1038/s41598-017-14945-1
34. Kumar R.D., Searleman A.C., Swamidass S.J., Griffith O.L., Bose R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics*, 2015, vol. 31, no. 22, pp. 3561–3568. https://doi.org/10.1093/bioinformatics/btv430
35. Akmal M.A., Hussain W., Rasool N., Khan Y.D., Khan S.A., Chou K.-C. Using CHOU'S 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, vol. 18, no. 5, pp. 2045–2056. https://doi.org/10.1109/TCBB.2020.2968441
36. Khan Y.D., Ahmad F., Anwar M.W. Aneuro-cognitive approach for iris recognition using back propagation. *World Applied Sciences Journal*, 2012, vol. 16, no. 5, pp. 678–685.
37. Khan Y.D., Ahmed F., Khan S.A. Situation recognition using image moments and recurrent neural networks. *Neural Computing and Applications*, 2014, vol. 24, no. 7–8, pp. 1519–1529. https://doi.org/10.1007/s00521-013-1372-4
38. Khan Y.D., Khan N.S., Farooq S., Abid A., Khan S.A., Ahmad F., Mahmood M.K. An efficient algorithm for recognition of human actions. *Scientific World Journal*, 2014, vol. 2014, pp. 875879. https://doi.org/10.1155/2014/875879
39. Khan Y.D., Khan S.A., Ahmad F., Islam S. Iris recognition using image moments and K-means algorithm. *Scientific World Journal*, 2014, vol. 2014, pp. 723595. https://doi.org/10.1155/2014/723595
40. Mahmood S., Khan Y.D., Mahmood M.K. A treatise to vision enhancement and color fusion techniques in night vision devices. *Multimedia Tools and Applications*, 2018, vol. 77, no. 2, pp. 2689–2737. https://doi.org/10.1007/s11042-017-4365-y
41. Butt H., Rasool N., Khan Y.D. A treatise to computational approaches towards prediction of membrane protein and its subtypes. *The Journal of Membrane Biology*, 2017, vol. 250, no. 1, pp. 55–76. https://doi.org/10.1007/s00232-016-9937-7
42. Akmal M.A., Rasool N., Khan Y.D. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS ONE*, 2017, vol. 12, no. 8, pp. 1–21. https://doi.org/10.1371/journal.pone.0181966
43. Pundir S., Magrane M., Martin M.J., O'Donovan C. Searching and navigating UniProt databases. *Current Protocols in Bioinformatics*, 2015. pp. 1.27.1–1.27.10 https://doi.org/10.1002/0471250953.bi0127s50
44. Delorenzi M., Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, 2002, vol. 18, no. 4, pp. 617–625. https://doi.org/10.1093/bioinformatics/18.4.617
45. Jia J., Liu Z., Xiao X., Liu B., Chou K.-C. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Analytical Biochemistry*, 2016, vol. 497, pp. 48–56. https://doi.org/10.1016/j.ab.2015.12.009

incorporating nucleotide physicochemical properties into PseKNC // Genomics. 2018. V. 111. N 1. P. 96–102. https://doi.org/10.1016/j.ygeno.2018.01.005
31. Huang C.H., Peng H.S., Ng K.L. Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms // BioMed Research International. 2015. V. 2015. P. 312047. https://doi.org/10.1155/2015/312047
32. Rahman M.S., Shatabda S., Saha S., Kaykobad M., Rahman M.S. DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC // Journal of Theoretical Biology. 2018. V. 452. P. 22–34. https://doi.org/10.1016/j.jtbi.2018.05.006
33. Chowdhury S.Y., Shatabda S., Dehzangi A. iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features // Scientific Reports. 2017. V. 7. P. 14938. https://doi.org/10.1038/s41598-017-14945-1
34. Kumar R.D., Searleman A.C., Swamidass S.J., Griffith O.L., Bose R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data // Bioinformatics. 2015. V. 31. N 22. P. 3561–3568. https://doi.org/10.1093/bioinformatics/btv430
35. Akmal M.A., Hussain W., Rasool N., Khan Y.D., Khan S.A., Chou K.-C. Using CHOU'S 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021. V. 18. N 5. P. 2045–2056. https://doi.org/10.1109/TCBB.2020.2968441
36. Khan Y.D., Ahmad F., Anwar M.W. Aneuro-cognitive approach for iris recognition using back propagation // World Applied Sciences Journal. 2012. V. 16. N 5. P. 678–685.
37. Khan Y.D., Ahmed F., Khan S.A. Situation recognition using image moments and recurrent neural networks // Neural Computing and Applications. 2014. V. 24. N 7–8. P. 1519–1529. https://doi.org/10.1007/s00521-013-1372-4
38. Khan Y.D., Khan N.S., Farooq S., Abid A., Khan S.A., Ahmad F., Mahmood M.K. An efficient algorithm for recognition of human actions // Scientific World Journal. 2014. V. 2014. P. 875879. https://doi.org/10.1155/2014/875879
39. Khan Y.D., Khan S.A., Ahmad F., Islam S. Iris recognition using image moments and K-means algorithm // Scientific World Journal. 2014. V. 2014. P. 723595. https://doi.org/10.1155/2014/723595
40. Mahmood S., Khan Y.D., Mahmood M.K. A treatise to vision enhancement and color fusion techniques in night vision devices // Multimedia Tools and Applications. 2018. V. 77. N 2. P. 2689–2737. https://doi.org/10.1007/s11042-017-4365-y
41. Butt H., Rasool N., Khan Y.D. A treatise to computational approaches towards prediction of membrane protein and its subtypes // The Journal of Membrane Biology. 2017. V. 250. N 1. P. 55–76. https://doi.org/10.1007/s00232-016-9937-7
42. Akmal M.A., Rasool N., Khan Y.D. Prediction of N-linked glycosylation sites using position relative features and statistical moments // PLoS ONE. 2017. V. 12. N 8. P. 1–21. https://doi.org/10.1371/journal.pone.0181966
43. Pundir S., Magrane M., Martin M.J., O'Donovan C. Searching and navigating UniProt databases // Current Protocols in Bioinformatics. 2015. P. 1.27.1–1.27.10 https://doi.org/10.1002/0471250953.bi0127s50
44. Delorenzi M., Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions // Bioinformatics. 2002. V. 18. N 4. P. 617–625. https://doi.org/10.1093/bioinformatics/18.4.617
45. Jia J., Liu Z., Xiao X., Liu B., Chou K.-C. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset // Analytical Biochemistry. 2016. V. 497. P. 48–56. https://doi.org/10.1016/j.ab.2015.12.009

Authors

Manickam Vijayalakshmi — Research Scholar, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Assistant Professor, Sri Sarada College for Women, Tirunelveli, 627011, India, https://orcid.org/0009-0001-2012-3169, vijimarthresearch@gmail.com
Mahesh Vallinayagi — PhD, Head, Associate Professor, Sri Sarada College for Women, Tirunelveli, 627011, India, https://orcid.org/0009-0006-0552-0138, vallinayagimahesh@gmail.com

Received 27.09.2023
Approved after reviewing 04.01.2024
Accepted 30.01.2024

Авторы

Виджаялакшми Маникам — научный сотрудник, Университет Манонманиам Сундаранар, Абишекапати, Тируневелли-627012; доцент, Женский колледж Шри Сарада, Тирунелвели, 627011, Индия, https://orcid.org/0009-0001-2012-3169, vijimarthresearch@gmail.com
Валлинаяги Махеш — PhD, руководитель, доцент, Женский колледж Шри Сарада, Тирунелвели, 627011, Индия, https://orcid.org/0009-0006-0552-0138, vallinayagimahesh@gmail.com

Статья поступила в редакцию 27.09.2023
Одобрена после рецензирования 04.01.2024
Принята к печати 30.01.2024