MPPIF-Net: Identification of Plasmodium Falciparum Parasite Mitochondrial Proteins Using Deep Features with Multilayer Bi-directional LSTM

Authors:

Samee Ullah Khan, Ran Baik

Date Submitted: 2020-09-23

Keywords: bi-directional LSTM, plasmodium falciparum, Machine Learning, mitochondrial protein

Abstract:

Mitochondrial proteins of Plasmodium falciparum (MPPF) are an important target for anti-malarial drugs, but their identification through manual experimentation is costly, and in turn, their related drugs production by pharmaceutical institutions involves a prolonged time duration. Therefore, it is highly desirable for pharmaceutical companies to develop computationally automated and reliable approach to identify proteins precisely, resulting in appropriate drug production in a timely manner. In this direction, several computationally intelligent techniques are developed to extract local features from biological sequences using machine learning methods followed by various classifiers to discriminate the nature of proteins. Unfortunately, these techniques demonstrate poor performance while capturing contextual features from sequence patterns, yielding non-representative classifiers. In this paper, we proposed a sequence-based framework to extract deep and representative features that are trust-worthy for Plasmodium mitochondrial proteins identification. The backbone of the proposed framework is MPPF identification-net (MPPFI-Net), that is based on a convolutional neural network (CNN) with multilayer bi-directional long short-term memory (MBD-LSTM). MPPIF-Net inputs protein sequences, passes through various convolution and pooling layers to optimally extract learned features. We pass these features into our sequence learning mechanism, MBD-LSTM, that is particularly trained to classify them into their relevant classes. Our proposed model is experimentally evaluated on newly prepared dataset PF2095 and two existing benchmark datasets i.e., PF175 and MPD using the holdout method. The proposed method achieved 97.6%, 97.1%, and 99.5% testing accuracy on PF2095, PF175, and MPD datasets, respectively, which outperformed the state-of-the-art approaches.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2020.1000
Citation (this specific file, latest version):	LAPSE:2020.1000-1
Citation (this specific file, this version):	LAPSE:2020.1000-1v1

DOI of Published Version: https://doi.org/10.3390/pr8060725

processes

MDPI

Article

MPPIF-Net: Identification of Plasmodium Falciparum Parasite Mitochondrial Proteins Using Deep Features with Multilayer Bi-directional LSTM

Samee Ullah Khan [1] and Ran Baik [2,*]

[1] Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Korea; sameek3797@gmail.com

[2] Department of Computer Engineering, Convergence School of ICT, Honam University, #417 Eodeung-daero, Gwangsan-gu, Gwangju 506-090, Korea

* Correspondence: baik@honam.ac.kr or ranbaik@gmail.com

Received: 29 April 2020; Accepted: 15 June 2020; Published: 22 June 2020

Abstract: Mitochondrial proteins of Plasmodium falciparum (MPPF) are an important target for anti-malarial drugs, but their identification through manual experimentation is costly, and in turn, their related drugs production by pharmaceutical institutions involves a prolonged time duration. Therefore, it is highly desirable for pharmaceutical companies to develop computationally automated and reliable approach to identify proteins precisely, resulting in appropriate drug production in a timely manner. In this direction, several computationally intelligent techniques are developed to extract local features from biological sequences using machine learning methods followed by various classifiers to discriminate the nature of proteins. Unfortunately, these techniques demonstrate poor performance while capturing contextual features from sequence patterns, yielding non-representative classifiers. In this paper, we proposed a sequence-based framework to extract deep and representative features that are trust-worthy for Plasmodium mitochondrial proteins identification. The backbone of the proposed framework is MPPF identification-net (MPPFI-Net), that is based on a convolutional neural network (CNN) with multilayer bi-directional long short-term memory (MBD-LSTM). MPPIF-Net inputs protein sequences, passes through various convolution and pooling layers to optimally extract learned features. We pass these features into our sequence learning mechanism, MBD-LSTM, that is particularly trained to classify them into their relevant classes. Our proposed model is experimentally evaluated on newly prepared dataset PF2095 and two existing benchmark datasets i.e., PF175 and MPD using the holdout method. The proposed method achieved 97.6%, 97.1%, and 99.5% testing accuracy on PF2095, PF175, and MPD datasets, respectively, which outperformed the state-of-the-art approaches.

Keywords: mitochondrial protein; machine learning; bi-directional LSTM; plasmodium falciparum

1. Introduction

Plasmodium falciparum are a unicellular protozoan organisms and toxic species that cause malaria in humans. It degrades hemoglobin in the acidic environment provided by the food vacuole [1]. When a female anopheles mosquito attacks human, malaria infection begins in the form of sporozoites into the bloodstream and its life cycle adopts many different stages [2]. These sporozoites are then quickly passed into the human liver where they upsurge exponentially into their cells to form merozoites. Merozoites attack red blood cells (erythrocytes) and again multiply until the cells burst to become trophozoites, schizonts, and gametocytes, during the last three stages of the anopheles life cycle as shown in Figure 1.

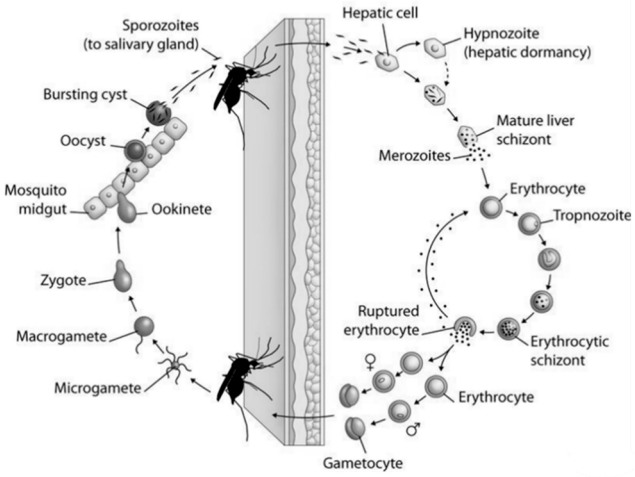

Figure 1. When a mosquito bites a human (host for malarial parasite) it causes infection by injecting sporozoites into the body, where it adversely affects hepatocyte's shape. Sporozoites grow rapidly in hepatocytes to become merozoites, while merozoites grow rapidly causing hepatocytes to burst and infect neighboring hepatocytes. When a mosquito bites the malaria patient the gametocytes that are produced from merozoites are taken by a mosquito. For the next 10 to 14 days the gametocytes produce sporozoites which are transferred to the saliva gland waiting for a mosquito to bite a healthy person and cause infection.

In eukaryotic cells, a mitochondrion is a membrane-bound organelle found in the cytoplasm which acts as a powerhouse of the cell responsible for cellular respiration and the production of adenosine triphosphate [3]. The inner membrane of cytoplasm comprises of different proteins such as enzymes, which are required for biochemical reactions. The mitochondrion has its own DNA (deoxyribonucleic acid) and ribosomes, which are 70 percent as that of prokaryote cells. In a cell, the mitochondrion is one of the important organelles which controls cellular metabolism and produces energy. Biologists have revealed that there are no significant similarities between mitochondrial proteins and human homologs [4].

Considering the constructive role of mitochondrial protein sequences in bioinformatics, proteomics, and cellular biology, many researcher's interest has been redirected to identify these biological sequences, but still it has been a challenging problem for them. With the invention of modern sequencing technologies, the number of these proteins has increased with rapid acceleration in the protein databanks. In 1990, only 3939 protein sequences were reported in the Uniprot database. According to the recent release statistics of protein databank, this number reached 550,000 in 2019 (11 December) [5].

There are two main approaches followed by researchers in the protein sequence prediction domain. The first category include machine learning-based approaches, where they employ features extractions methods in order to extract various patterns from biological sequences. Most of the existing literatures followed this approach. The second one is deep learning-based approach which extracts deep features and contextual information from proteins, which improves the prediction accuracy significantly. We briefly discuss the related works for both the mentioned approaches.

1.1. Machine Learning Approach towards Mitochondria Proteins Identification

In past decades, numerous machine learning algorithms and computational biological techniques are proposed for the categorization of mitochondrial and non-mitochondrial proteins via complex sequences. Bender et al. [6] evaluated MPPF by principal component analysis, statistical methods, and supervised neural network. They developed a model PlasMit based on extracting new composition patterns from proteins which efficiently predicted the mitochondrial proteins. R Verma et al. [7] combined two feature descriptors i.e., split amino acid and position specific scoring metrics, in

order to predict mitochondrial proteins accurately. Jia et al. [8] considered mitochondrial proteins as attractive targets for anti-malarial drugs, but manual identification of these proteins is a difficult and time-consuming task. Therefore, they used two proteins encoding approaches such as bi-profile Bayes and split amino acid composition in order to extract specific pattern features from the amino acid chain. Authors trained a support vector machine classifier on these statistical features for final prediction. Afridi et al. [9] proposed genetic programming and an ensemble approach based on the feature extraction method for mitochondrial protein classification. Ding et al. [10] analyzed the variance for the accurate prediction of mitochondrial proteins. They suggested that combining more and more features is not a reliable approach because it takes more time in execution and contained redundant values which degraded the performance of the model. Therefore, to reduce the dimensionality and select the optimal features, researchers in this article used analysis of variance (ANOVA). Due to the complexity of the Plasmodium falciparum genome, the prediction of MPPF is more difficult than other species. Chen et al. [11] proposed an n-peptide composition of reduced amino acid alphabet which is obtained from a structural alphabet named protein as a feature parameter; the increment of diversity was firstly proposed to predict mitochondrial proteins. For instance, Cai et al. [12] applied support vector machine (SVM) based algorithm to train a predictor on three types of feature descriptors technique including transition (T), composition (C), and distribution (D) for obtaining additional physicochemical properties of different amino acids. R Kumar et al. [13] proposed a two-level model named as SubMitoPred. In the first level, authors predicted mitochondrial proteins while in the second level, they forecasted the sub-classes of mitochondrial localization. The whole model was based on the combination of SVM and the Pfam information domain. For further improvement, C Savojardo et al. [14] developed the deep learning model DeepMito. They trained and tested the model on a new high-quality dataset. Furthermore, they also developed a webserver for predicting mitochondrial and sub-mitochondrial localization. DNA-binding proteins (DBP's) can be used for the regulation of transcription and gene expression along with the identification of particular nucleotides. Therefore, for accurate and precise predictions of DBP's, Waris et al. [15] used evolutionary profiles position-specific scoring matrix for sequence encoding and a support vector machine for classification. Most of the available drugs are prepared to target the membrane proteins. Discriminating these proteins via computer vision techniques is an effective and timesaving as well. Therefore, Hayat et al. [16] through efforts of a computational method, accurately predicted the membrane proteins via different machine learning algorithms.

1.2. Deep Learning Approach towards Mitochondria Proteins Identification

Some researchers have utilized the full benefits of deep learning techniques and employed them to predict different types of protein sequences [17–19]. Delong et al. [20] attempted for the first time, to generate the original idea in deep learning for the discrimination of DNA binding proteins and non-DNA binding proteins. Qinhu et al. [21] improved the inherent weak supervision biological information prediction by proposing a new procedure established on CNN features with sequence-based learning to classify the DNA binding proteins. Recently, for sequencing learning, Qu et al. [17] applied a sequence learning network known as a recurrent neural network (RNN) with CNNs to predict those proteins which are attached to DNA. Compared to the traditional methods, deep learning techniques enhance the flexibility of extracted optimal features from sequences. It is not only the selection of a large number of proteins that are made possible for model training, but the process of speedy and accurate prediction is also enhanced. From earlier scholar's work, in addition to protein features, contextual information has also been observed as a valuable feature [22]. With an inspiration of this concept, if an amino acid sequence also suppresses contextual features, then it might increase the prediction score.

In this article, a CNN model is proposed by employing MBD-LSTM for Plasmodium mitochondria protein identification. In the first encoding layer, we assign an integer value to each amino acid in order to encode protein sequences and find the maximum length of the protein, and the next layer is the embedded layer where each word is converted to fix the length vector. Further, these matrices

are passed through the CNN model containing three convolutional layers followed by max pooling layers. Finally, these deep features are fed into the MBD-LSTM layer for sequence learning and the last dense layer generates the optimal output. The proposed novel framework makes the following main contributions for the identification of Plasmodium falciparum parasite mitochondrial proteins.

- Considering the lack of effective vaccination, a rise in drug-resistant Plasmodium parasites, and the lethal nature of malaria, we propose a novel sequence-based framework MPPIF-Net to efficiently discriminate Plasmodium mitochondria and non-mitochondria proteins. The proposed model is useful in developing vaccines against malaria parasites.
- With the rising sequencing technology, the number of various proteins increases day by day with rapid acceleration in the protein databanks. In the aforementioned literature, researchers follow machine learning and computational techniques, which revealed inadequate performance while capturing contextual features from biological sequence patterns, yielding non-representative classifiers. In this study, we pursue a deep learning approach, which is capable of extracting contextual features and apply a sequence learning mechanism to efficiently classify the nature of proteins with the assistance of CNN and MBD-LSTM.
- Due to the unavailability of a large benchmark dataset of Plasmodium mitochondrial proteins, in this paper, we prepared a new dataset from the Uniprot site which contains both mitochondria and non-mitochondria proteins. The types of proteins mentioned in our dataset are passed from CD-Hit software to detect and remove similarity and short length proteins to optimally acquire a preprocessed and adoptable dataset.
- To validate the adoptability of our proposed model, we also made an extensive experimentation on the benchmark datasets, that is designed using mitochondrial proteins of another organism. The proposed model responded with convincing accuracy on this dataset, thereby validating the fact that our model is adoptable not only to the mitochondria proteins of the Plasmodium organism, but is trust-worthy to classify mitochondria proteins of other species as well.

The remaining article is divided into three further sections; Section 2 briefly explains the proposed system. Results and discussion are explained in Section 3 and in Section 4 we present the conclusions and future directions.

2. Proposed Methodology

Our proposed model mainly contains three modules which are further divided into five phases: (1) the preprocessing phase, in which we collect the sequence data from the protein databank and apply some techniques for the refining of data; (2) the encoding phase, in which we simply assign a natural number to each amino acid and also search the maximum length of the sequence; (3) the embedding phase, which is a mapping procedure in which individual words in the separate vocabulary will be inserted into a continuous vector space; (4) the convolution phase, in which we create one-dimensional vector from the encoded amino acid that is advanced to the CNN layers for deep features extraction; and finally, (5) the MBD-LSTM phase, where all the features are passed through this network for sequence learning to generate final output, as shown in Figure 2. The details of all phases are given in the subsequent sections. After passing the protein sequence (Pseq), the affinity scores of mitochondrial proteins are calculated by the Equation (1).

$$A(Pseq) = (Encoding\ (Embedding\ (CNN\ (MBD\text{-}LSTM))) \quad (1)$$

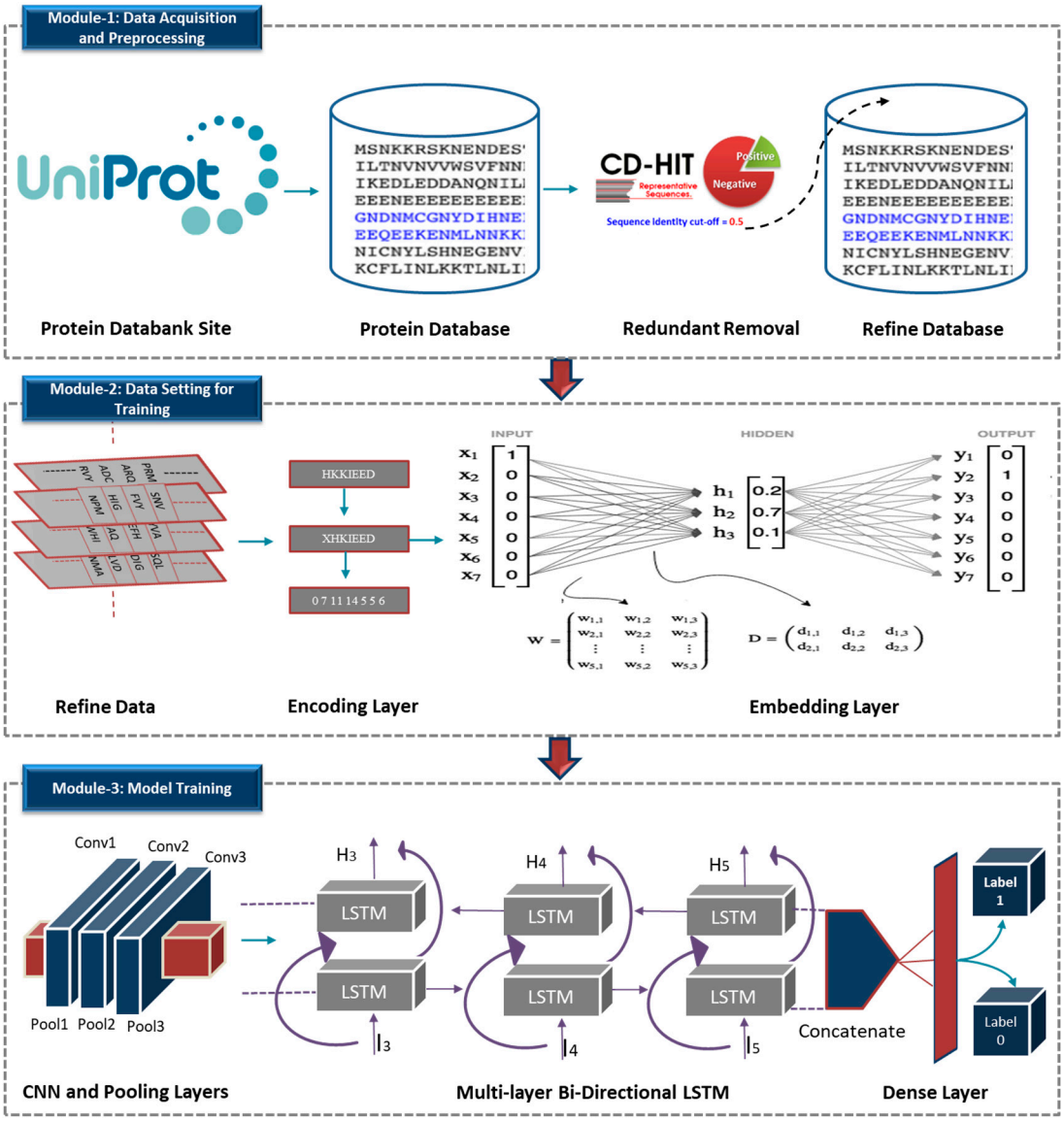

Figure 2. The proposed framework comprises of three modules. Module 1 describes the phenomena of sequence acquisition in which collection and preprocessing is channeled to eliminate redundancy. The polished sequences are forwarded to Module 2, where the alphabet was converted to natural numbers; after that we utilized embedding layers to generate fixed length vectors. Finally, in Module 3, we passed one-dimensional data to CNN deep contextual features extraction and then employed multi-layer bi-directional long short-term memory MBD-LSTM for sequence learning. Afterword, sigmoid activation is applied to predict final probability scores; either the output is related to mitochondrial (Label 1) or non-mitochondria (Label 0) which are then evaluated in terms of accuracy.

For predicting the output label data, we used the sigmoid activation function, and to evaluate the performance of the network, binary cross entropy is applied.

2.1. Raw Data Acquisition and Preprocessing

The biological sequences are obtained from the Uniprot protein databank in FASTA format, which is basically a text-based format. It is easily accessible, and downloadable protein sequences of any organism. In this study, we utilized three datasets such as PF2095, PF175, and MPD to estimate the model accuracy. Due to the unavailability of the massive number of MPPF, we collected 1701 raw sequences by searching the keywords "plasmodium falciparum mitochondrion", that are considered as

positive samples. Furthermore, we collected 2075 negative samples by searching the "non-mitochondria proteins". After extracting the sequences, we inserted these sequences to CD-Hit software to removed sequences with 80% similarity and shorter length (less than 40) of amino acids. After the refining process, we collected 890 and 1205 as positive and negative samples for classification purposes. The remaining two benchmark datasets were publicly available, so there is no need to pass it from the preprocessing phase.

2.2. Encoding Protein Sequences

In machine learning, various feature extraction approaches are proposed, such as amino acid composition, dipeptide composition, split amino acid composition, pseudo amino acid composition, position specific scoring matrices, and n-gram methods, etc. [18]. These manual protein encoding techniques usually extract low-level features which sometimes degrade the machine learning model, mostly in classification problems. Nowadays, a renowned approach for the achievement of a high success rate is deep learning mechanism, which is the subset of machine learning in artificial intelligence. Here we simply allocate a natural number to each amino acid [23]. For instance, we have a protein sequence like 'AKILMEF', so the encoding of each amino acid is represented as (1, 4, 5, 8, 9, 10, 11). Symbol along with code for each amino acid is shown in Table 1. This encoding is efficient as compared to the sparse vector. An important aspect for consideration while assigning numbers to amino acids is related to order, which states that order does not affect a model's performance at all.

Table 1. Symbolic representation of each Amino Acids and its code.

Amino Acids	Letters	Code
Alanine	A	1
Cysteine	C	2
Aspartic	D	3
Glutamic	E	4
Phenylalanine	F	5
Glycine	E	6
Histidine	H	7
Isoleucine	I	8
Lysine	K	9
Leucine	L	10
Methionine	M	11
Asparagine	N	12
Proline	P	13
Glutamine	Q	14
Arginine	R	15
Serine	S	16
Threonine	T	17
Valine	V	18
Tryptophan	W	19
Tyrosine	Y	20

The encoding phase only creates a digital vector of a proteins sequence with variable length. First, we find the maximum length of protein in a dataset. In this paper, the max-length vector value is (5253, 1280, and 1402) for distinct datasets which usually depends upon the sequences in the dataset. As we already know, protein sequences usually exhibit different lengths while in deep learning, we are required to keep a fixed length for all protein sequences. For example, for a protein whose length vector is smaller than the maximum-length vector, a unique value zero is placed, at the end in order to keep the same alignment of all the sequences. An encoding of protein example is given in Equation (2).

$$\text{Protein Seq1} = \text{Encoding (Sequence)} = (11, 12, 16, 17, 0) \quad (2)$$

2.3. Embedding Layer

It is a very difficult task to encode each word manually. This layer gives us an automatic and efficient way of representing words or documents in which matching words have a similar encoding [16]. This work is done by just multiplying one hot vector from the left with a weight matrix W $\epsilon R^{d} * |V|$ where |V| represents the number of primary symbols related to the vocabulary as shown in Equation (3).

$$V_Z = W_{X_t} \tag{3}$$

As a result, the input sequence of amino acids becomes a solid valued vector (z = 1, 2, 3, 4, … n). In the embedding layer, assume that the output dense vector length is 8, and each number map corresponds to a fixed vector length. After passing through layer proteins, the sequence becomes an 8×8 matrix e.g., as exposed in Equation (4). We may represent Thyronine amino acid with [0.5, −0.8, 0.7, 0.4, 0.3, −0.5, −0.7, 0.8] and Methionine with [0.4, −0.4, 0.5, 0.6, 0.2, −0.1, −0.3, 0.2].

$$\text{ProteinSeq2} = \begin{pmatrix} 0.1 & -0.4 & 0.1 & 0.2 & 0.6 & 0.4 & -0.1 & 0.1 \\ 0.4 & -0.4 & 0.5 & 0.6 & 0.2 & -0.1 & -0.3 & 0.2 \\ 0.2 & -0.2 & 0.6 & 0.7 & -0.1 & 0.1 & -0.2 & 0.1 \\ 0.5 & -0.2 & 0.1 & 0.6 & 0.2 & -0.6 & -0.2 & 0.9 \\ 0.4 & -0.4 & 0.5 & 0.6 & 0.2 & -0.1 & -0.3 & 0.2 \\ 0.8 & -0.5 & 0.4 & 0.7 & 0.5 & -0.2 & -0.5 & 0.3 \\ 0.9 & -0.6 & 0.7 & 0.8 & 0.2 & -0.1 & -0.2 & 0.7 \\ 0.5 & -0.8 & 0.7 & 0.4 & 0.3 & -0.5 & -0.7 & 0.8 \end{pmatrix} \tag{4}$$

2.4. Convolution Layer

The deep learning model works very efficiently in image processing and video analysis by extracting the deep features based on convolutional and pooling layers [24,25]. In case of images or videos, we directly give input data to the model because their data already exhibits a matrix arrangement [26]. While working with protein sequences, first, we prepare data in the form of a matrix with fixed-size and forwards to the convolution layer for processing like images. In this work, the model exhibits three convolution layers and each one is followed by a max pooling layer for deep features extraction. In this layer, we use 3 × 8 filters to scan the protein seq2 and obtain a new feature map as shown in Equations (5) and (6).

$$\textbf{Filter} = \begin{bmatrix} 0.2 & 0.2 & -0.3 & 0.8 & 0.5 & 0.3 & 0.2 & -0.2 \\ 0.1 & 0.3 & -0.3 & 0.6 & 0.1 & 0.3 & -0.2 & 0.3 \\ 0.8 & -0.2 & 0.3 & -0.5 & 0.6 & 0.3 & 0.2 & 0.1 \end{bmatrix} \tag{5}$$

$$\text{Protein seq3} = \text{Convolution (Seq2)} = \begin{bmatrix} 0.48 \\ 0.53 \\ 0.75 \\ 0.20 \\ 0.25 \\ 0.62 \\ 0.40 \end{bmatrix} \tag{6}$$

In max pooling layer, the sliding window takes the highest value of the two numbers as shown in Equation (7)

$$\text{Protein seq4} = \text{Max Pooling (Seq3)} = \begin{bmatrix} 0.65 \\ 0.53 \\ 0.48 \\ 0.62 \end{bmatrix} \tag{7}$$

2.5. MBD-LSTM Layer

For the complications and issues related to short-term memory sequencing, RNN is employed, mostly for the cases where a long sequence is required to handle and stored for both forward and backward steps. As a result of the continuation of this procedure, RNN may depart crucial information out of the initial sequence data. But during backpropagation, a vanishing gradient issue is encountered, that makes it hard to memorize long-term changes in sequence [27,28]. Throughout the propagation process, the neural network weights are updated and shrink due to the gradient. These extremely minor weights do not participate to the learning process in an RNN, and also layers stop learning due to acquiring such a small gradient. In this situation the RNN does not have a capability to store longer sequence modifications that were observed previously. LSTM provides a solution by incorporating a short-term memory unit which is a special recurrent neural network architecture. LSTM emphases build memory cells and gates that regulate to process and store information and also allow when to update and forget the hidden states of the network [29]. The internal structure review of LSTM contains memory cell state $S_{st\text{-}1.}$ These cells directly relate to H_{st}-1 which is the middle output state, and the successive state X_{st} controls the internal state vector which is required to be upgraded. There are three gates in LSTM structure; input gates N_{st}, forget gates F_{st}, and the output gate O_{st}. The mathematical notation of these gates are as follows.

$$F_{ST} = \sigma(W_{FX}X_{ST} + W_{FH}\, H_{ST-1} + B_F) \tag{8}$$

$$I_{ST} = \sigma(W_{IX}X_{ST} + W_{IH}\, H_{ST-1} + B_I) \tag{9}$$

$$N_{ST} = \phi(W_{NX}X_{ST} + W_{NH}\, H_{ST-1} + B_N) \tag{10}$$

$$O_{ST} = \sigma(W_{OX}X_{ST} + W_{OH}\, H_{ST-1} + B_O) \tag{11}$$

$$S_{ST} = \sigma(G_T \theta I_{ST} + S_{ST-1}\, \theta\, B_S) \tag{12}$$

$$H_{ST} = \phi(S_{ST-1})\, \theta O_{ST} \tag{13}$$

In Equations (8)–(13), the network inputs weight matrices are represented by W_{FX}, W_{FH}, W_{IX}, W_{IH}, W_{NX}, W_{NH}, W_{OX}, and W_{HO}. Here, θ is used for the multiplication in an elementwise manner. The two activation functions such as sigmoid and tanh are represented by σ and ϕ. The single time step of LSTM architecture is shown in Figure 3a. In this article, we evaluated the performance of MBD-LSTM for protein sequence identification. The idea of a MBD-LSTM is developed from traditional bidirectional RNN [30], which also processes the hidden layer input sequence data in both forward and backward direction. MBD-LSTM has achieved significant results in speech recognition [31], summarization [32], classification, energy consumption prediction [33], and text generation. The structure of MBD-LSTM consists of forward and backward layers as shown in Figure 3b. The output of the forward layer $I_T^{>}$ is analyzed through input data from T − n to T − 1, while the output data of the backward layer $H_T^{<}$ is generated through reversed inputs such as from T − n to T − 1. Final MBD-LSTM generates the O_T output vector as illustrated in equation (14).

$$O_T = \sigma(I_T^{>},\, H_T^{>}) \tag{14}$$

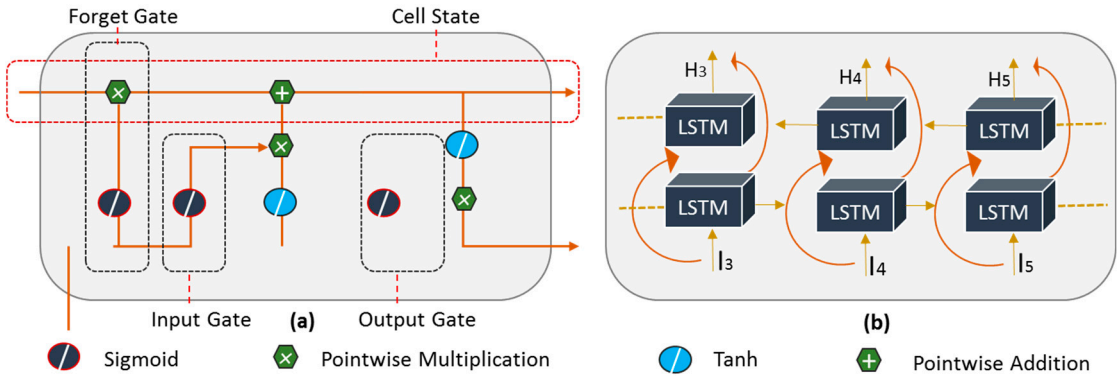

Figure 3. (**a**) Internal architecture of LSTM, comprising of multiple gates along with LSTM cells for stimulation of numerous operations, permitting the gates to store and omit related information; (**b**) MBD-LSTM, which acquires input data sequences and then proceeds in a forward and backward direction.

In Equation (11), σ combines an output sequence of two layers, which is also known as summation function.

3. Results and Discussion

In this portion, an in-depth analysis over comprehensive experiments which are performed on three protein sequence datasets and detailed discussion of comparative studies of the proposed model with state-of-the-art techniques is presented.

3.1. Datasets

The benchmark dataset of Plasmodium mitochondrial protein sequences was obtained from [5]. The total number of proteins in dataset is 175, which contains 40 positive (mitochondrial proteins) and 135 negative (non-mitochondrial proteins) samples. We represent this dataset by Plasmodium falciparum 175 (PF175). The second raw dataset (PF2095) was obtained from the universal protein resource (Uniprot) which contains 890 positive samples and 1205 negative samples.

Similarly, a third dataset was downloaded from [34] which contained 499 positive samples. In this paper, we denote this dataset by (MPD) as shown in Table 2. Actually, 2833 proteins were obtained from the protein databank site known as Swiss-Prot by searching the keyword mitochondrial. Afterward, those proteins were then excluded with ambiguous words, such as SIMILARITY, POTENTIAL, or PROBABLE and FRAGMENTS. Furthermore, 681 proteins were collected belonging to locations other than mitochondrial site.

Table 2. Plasmodium falciparum mitochondria protein datasets.

Dataset	Positive Sample	Negative Sample
PF175	40	135
PF2095	890	1205
MPD	499	250

By applying the preprocessing and eliminating the ambiguous data, we selected 250 proteins as non-mitochondrial.

3.2. Experimental Setup

Using two benchmark and one own prepared protein datasets, we analyzed and verified the efficiency of the proposed model. The model was trained on a Titan Intel Core i5-6600 processor with X (Pascal)/PCLe/SSE2 GPU, having 64GB of memory using 16.4 LTS Ubuntu operating system.

The proposed deep learning model was executed in version 3.5 of python, version 2.2.4 of Keras, and version 1.12 of TensorFlow backend along with an Adam employed as an optimizer. To find the most favorable selection of the hyperparameter of each model, several experiments were conducted. At last, we selected 50 epochs to train the model with a batch size of 100. The PF2095 and MPD samples were split into training 70% and testing 30%, and due to fewer numbers of protein samples in PF175 we kept 80% data in training, and 20% data is utilized for model evaluation.

3.3. Evaluation Metrics

In this study, a couple of assessment measures are used for the evaluation of the proposed model. These parameters include accuracy, sensitivity, and specificity. The mathematical formulas are defined in the Equation (15)–(17).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{15}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{16}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{17}$$

Now, let us assume that the mitochondrial protein is positive, and the non-mitochondrial protein is negative. The true positive (TP) is that value in which predictive and actual value is positive and true negative (TN) is the value in which predicted value and actual value is negative. Similarly, the false positive (FP) is the value in which a machine predicted as positive but actually it is a negative value and the false negative (FN) is that value in which machine predicted as negative class but actually it is related to positive class value.

3.4. Ablation Study on PF2095

In this subsection, we conduct an ablation study after comprehensive experiments to analyze the three models in terms of accuracy, sensitivity, and specificity on the PF2095 dataset, which is a new mitochondria proteins of Plasmodium dataset comprising 890 positive and 1205 negative samples. In this dataset 70% of total samples are set for training, and the remaining 30% are used for model evaluation. First, we perform our experiments on CNN-GRU which achieved 89.7% training accuracy, 88.0% testing, 90.4% sensitivity, and 88.9% specificity. The next model CNN-LSTM showed better performance compared to previous one. It obtained 93.5% training, 91.2% testing accuracy, 90.6% sensitivity, and 91.7% specificity. The proposed model MPPIF-NET used CNN with integration of MBD-LSTM with the same number of parameters. It is experimentally proved that the last hybrid approach shows supremacy of performance which obtained 98.2% training accuracy, 97.6% is testing performance of the model, 98.1% of its sensitivity, and 97.2% specificity. The detailed experimental evaluation results are depicted in Table 3 and confusion metrics of the MPPIF-NET are shown in Figure 4.

Table 3. Training and testing performance of the MPPIF-NET on different models and datasets.

Dataset	Model	Training Accuracy	Testing Accuracy	Sensitivity	Specificity
PF2095	CNN-GRU	89.7	88.0	90.4	88.9
	CNN-LSTM	93.5	91.2	90.6	91.7
	CNN-MBD-LSTM (Proposed)	98.2	97.6	98.1	97.2
PF175	CNN-MBD-LSTM (Proposed)	100	97.1	100	96.2
MPD	CNN-MBD-LSTM (Proposed)	99.7	99.5	99.3	100

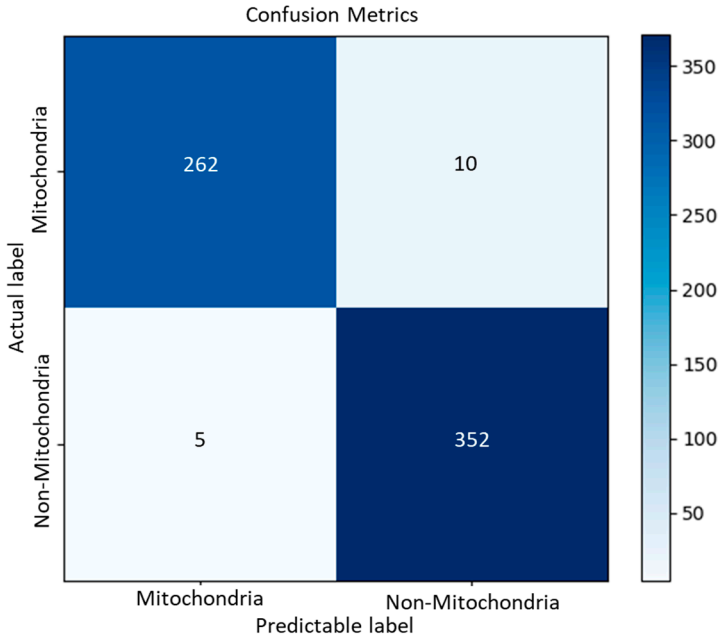

Figure 4. Confusion metrics of the proposed method over the PF2095 dataset.

3.5. Experimental Evaluation on PF175

We used irregular data in our experiments along with a hold-out technique which is the simplest kind of cross validation. The data was divided into training and testing. We trained our proposed model on 80% of the data and the remaining 20% of data were used for evaluation purposes. During experiments we updated different parameters to achieve good performance. After numerous experiments we set these parameters and their value; for example, maximum length of the protein is 1280 which depends upon the dataset, maximum features = 26, embedding size = 8, number of filters in convolutional are 32, pooling length = 2, batch size = 100, dropout = 0.2, and number of epochs is 50. We also checked different numbers of epochs and finally realized that the trained model fits the protein sequences well and predicts accurately on epochs 50.

Our model achieved better performance in terms of 100% training accuracy, 100% sensitivity, 96.2% specificity, and testing accuracy of 97.14%, which is higher than other state-of-the-art approaches. The confusion matrics of correctly and incorrectly predicted proteins are shown in Figure 5.

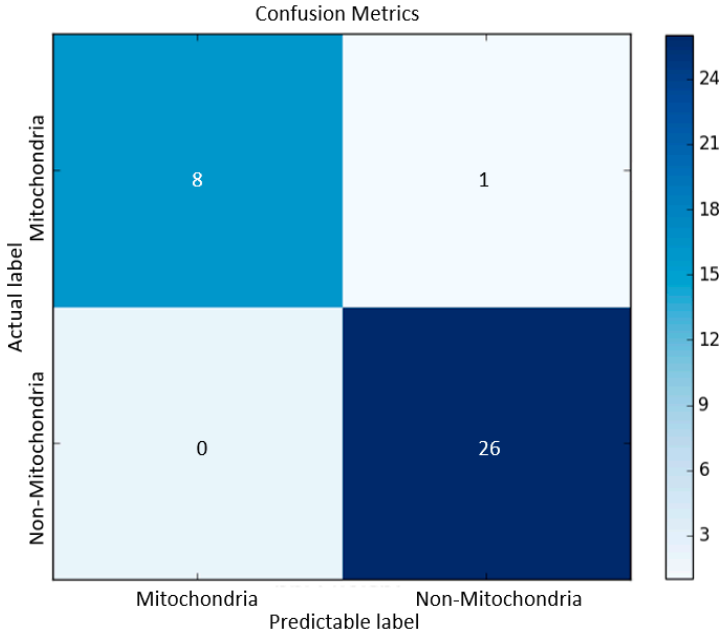

Figure 5. Confusion metrics of the proposed method over the PF175 dataset.

3.6. Experimental Evaluation on MPD

This dataset is also an unbalanced dataset and used the hold-out method during experiments. The data is divided into training and testing, which is 70% and 30%. For this dataset we also set the same parameters, except the maximum length of the protein which is 1402; maximum features = 26, embedding size = 8, number of filters in convolutional are 32, pooling length = 2, batch size = 100, dropout = 0.2 and the number of epochs is 50. We have done a lot of experiments with different setup parameters, but finally on epochs 50 we achieved better performance. The confusion matrics of correctly and incorrectly predicted proteins are shown in Figure 6.

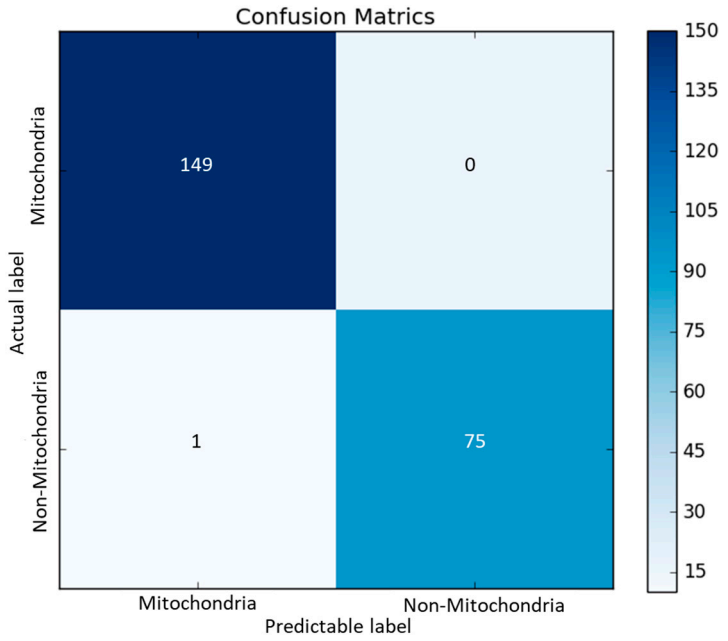

Figure 6. Confusion metrics of the proposed method over the mitochondria protein dataset (MPD) dataset.

Our model got 99.7% training accuracy, 99.3% sensitivity, 100% specificity, and a testing accuracy of 99.5%, which shows that the proposed model is superior in contrast to state-of-the-art techniques.

3.7. Comparative Analysis of the MPPIF-NET with Other Models on PF175

In the post genomic era, functional annotation is one of the major challenges. From the last decade, a vast number of machine learning and bioinformatic techniques have been proposed to predict protein functionality. The statistics of sequences are boosted day by day in the protein databanks. Identification of these biological sequences via laboratory methods was a laborious task. Therefore, we proposed a deep learning model for the accurate prediction of a huge number of proteins. Hence, it is important to evaluate the performance of models in order to compute the realistic performance of the model. For this we compared our proposed model with the state-of-the-art method using the same dataset. In the first attempt Bhasin et al. [35] proposed a model for eukaryotic subcellular localization protein prediction called (Eslpred) using a hybrid approach containing a dipeptide composition and PSI-BLAST. They achieved 69.71% accuracy, 73.33% specificity, and 57.50% sensitivity. Guda et al. [36] developed a new method for genome-scale prediction of the target mitochondria protein based on the composition of the amino acid and the occurrence frequency of each pattern which repeats in sequences. They achieved 80% accuracy, 87.41% specificity, and 55% sensitivity. Bender et al. [6] built a neural network model for the precise prediction of mitochondrial transit peptides which causes malaria. Due to the complex genomic sequence of PF, Chen et al. [11] developed the increment of diversity model in which a reduced amino acid composition was used in order to extract local features from the biological sequence. The prediction performance achieve 100% superior sensitivity rate, 89% specificity, and 92% accuracy as shown in Table (4). Mitochondria are vital organelles of eukaryotic cells which are involved in processing cellular death and human diseases; therefore, Afridi et al. [9] proposed an ensemble model known as Mito-GSAAC in which the main purpose was to examine an effective feature extraction approach. They achieved the highest specificity score of 95.56%, 93.21% accuracy, and 87.5% sensitivity. Accurate identification of the mitochondrial protein of Plasmodium falciparum is an essential role in the discovery of anti-malarial drug targets. Ding et al. [10] used a dipeptide composition for protein encoding. They also used the analysis of variance to overcome the issue of overfitting. They attained 97.1% accuracy, 90% sensitivity, and 99.3% specificity. The aforementioned state-of-the-art techniques utilized the machine learning approaches for the protein sequences prediction. We proposed a deep learning strategy for identification of these biological sequences which gave 97.14% superior testing accuracy compared to other discussed methods as shown in Table 4.

Table 4. MPPIF-NET comparative analysis with other models on PF175 dataset.

Method	Sensitivity	Specificity	Accuracy
Eslpred [34]	57.50	73.33	69.71
Mitopred [35]	55.00	87.41	80.00
PlasMit [5]	94.00	89.00	90.00
ID [10]	100	89.63	92.00
MitoGSAAC [8]	87.5	95.56	93.21
ANOVA [9]	90.0	**99.3**	97.1
MPPFI-Net	**100**	96.2	**97.14**

3.8. Comparative Analysis of the MPPIF-NET with Other Models on MPD

Mitochondria are the center and powerhouse of the eukaryotic cells. Pharmaceutical companies still desire such a system which accurately predicts the mitochondria protein of Plasmodium in order to prepare drugs. Therefore, Tan et al. [34] proposed an algorithm in order to evaluate the pair composition of amino acids. The extracted features are then passed to the support vector machine classifier for prediction of Plasmodium mitochondria proteins. The SVM model was evaluated which achieved 85% accuracy, 89.28% specificity, and 79.16% sensitivity. Jiang et al. [37] developed a new sequence-based

method which is known as the Discrete Wavelet Transform for sequence prediction. They achieved 50.30% sensitivity, 95.74% specificity, and 76.53% accuracy. Afridi et al. [9] used four computational methods such as AAC, DPC, SAAC, and PAAC. Furthermore, they also evaluated the six machine learning algorithms, such as support vector machine, random forest, multilayer perceptron, AdaBoost, and bagging. Finally, on the basis of the ensemble classifier they achieved 92.62% accuracy, 91.52% specificity, and 90.96% sensitivity. Our proposed model performs well compared to the state-of-the-art methods, having 99.5% accuracy, 100% specificity, and 99.33% sensitivity as shown in Table 5.

Table 5. MPPIF-NET comparative analysis with other models on the MPD dataset.

Method	Sensitivity	Specificity	Accuracy
SVM 84-D [33]	79.16	89.28	85.00
DWT [36]	50.30	95.74	76.53
MitoGSAAC [8]	90.96	91.52	92.62
MPPFI-Net	**100**	**99.33**	**99.5**

4. Conclusions and Future Directions

For the identification of mitochondria proteins of Plasmodium some biologists are still concentrating on extracting new patterns from biological sequences and are searching for appropriate machine learning algorithms which accurately classify proteins. In this study, we proposed a deep leaning framework MPPFI-Net which is capable of extracting deep features automatically and can discriminate proteins quickly and accurately. We merged the CNN and MBD-LSTM in order to extract the contextual information from amino acids. Later on, we compared MPPFI-Net performance with the state-of-the-art models, and we conclude that the proposed framework speeds up the performance regarding both prediction accuracy and fitting uncharacterized data. In future, we will boost this work by fusing the traditional features and deep features.

Author Contributions: Conceptualization, S.U.K. and R.B.; methodology, S.U.K.; writing—original draft preparation, S.U.K.; review and editing, S.U.K. and R.B.; supervision, R.B. All authors have read and agreed to the published version of the manuscript.

Funding: The publication fees were supported by Prof. Ran Baik (HONAM University-20190125).

Acknowledgments: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this study was supported by the research fund from Honam University (2019).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gazanion, E.; Vergnes, B. Protozoan parasite auxotrophies and metabolic dependencies. In *Metabolic Interaction in Infection*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 351–375.
2. Dundas, K.; Shears, M.J.; Sinnis, P.; Wright, G.J. Important extracellular interactions between Plasmodium sporozoites and host cells required for infection. *Trends Parasitol.* **2019**, *35*, 129–139. [CrossRef] [PubMed]
3. Hou, X.S.; Wang, H.S.; Mugaka, B.P.; Yang, G.J.; Ding, Y. Mitochondria: Promising organelle targets for cancer diagnosis and treatment. *Biomater. Sci.* **2018**, *6*, 2786–2797. [CrossRef]
4. Devine, M.J.; Kittler, J.T. Mitochondria at the neuronal presynapse in health and disease. *Nat. Rev. Neurosci.* **2018**, *19*, 63. [CrossRef] [PubMed]
5. UniProtKB/Swiss-Prot UniProt 2019. Available online: https://www.uniprot.org/statistics/Swiss-Prot%202019_06 (accessed on 20 May 2020).
6. Bender, A.; van Dooren, G.G.; Ralph, S.A.; McFadden, G.I.; Schneider, G. Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. *Mol. Biochem. Parasitol.* **2003**, *132*, 59–66. [CrossRef] [PubMed]
7. Verma, R.; Varshney, G.C.; Raghava, G.P.S. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. *Amino Acids* **2010**, *39*, 101–110.

8. Jia, C.; Liu, T.; Chang, A.K.; Zhai, Y. Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. *Biochimie* **2011**, *93*, 778–782. [CrossRef] [PubMed]
9. Afridi, T.H.; Khan, A.; Lee, Y.S. Mito-GSAAC: Mitochondria prediction using genetic ensemble classifier and split amino acid composition. *Amino Acids* **2012**, *42*, 1443–1454. [CrossRef] [PubMed]
10. Ding, H.; Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* **2015**, *47*, 329–333. [CrossRef] [PubMed]
11. Chen, Y.-L.; Li, Q.-Z.; Zhang, L.-Q. Using increment of diversity to predict mitochondrial proteins of malaria parasite: Integrating pseudo-amino acid composition and structural alphabet. *Amino Acids* **2012**, *42*, 1309–1316. [CrossRef] [PubMed]
12. Cai, C.Z.; Han, L.Y.; Ji, Z.L.; Chen, X.; Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003**, *31*, 3692–3697. [CrossRef]
13. Kumar, R.; Kumari, B.; Kumar, M. Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. *Mitochondrion* **2018**, *42*, 11–22. [CrossRef] [PubMed]
14. Savojardo, C.; Bruciaferri, N.; Tartari, G.; Martelli, P.L.; Casadio, R. DeepMito: Accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. *Bioinformatics* **2020**, *36*, 56–64. [CrossRef] [PubMed]
15. Waris, M.; Ahmad, K.; Kabir, M.; Hayat, M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* **2016**, *199*, 154–162. [CrossRef]
16. Hayat, M.; Khan, A. MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM. *J. Theor. Biol.* **2012**, *292*, 93–102.
17. Qu, Y.H.; Yu, H.; Gong, X.J.; Xu, J.H.; Lee, H.S. On the prediction of DNA-binding proteins only from primary sequences: A deep learning approach. *PLoS ONE* **2017**, *12*, e0188129. [CrossRef]
18. Zeng, H.; Edwards, M.D.; Liu, G.; Gifford, D.K. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* **2016**, *32*, i121–i127. [CrossRef] [PubMed]
19. Qiu, W.; Li, S.; Cui, X.; Yu, Z.; Wang, M.; Du, J.; Peng, Y.; Yu, B. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. *J. Theor. Boil.* **2018**, *450*, 86–103. [CrossRef] [PubMed]
20. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef] [PubMed]
21. Zhang, Q.; Zhu, L.; Bao, W.; Huang, D.S. Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**. [CrossRef]
22. Melamud, O.; Goldberger, J.; Dagan, I. context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 7–12 August 2016.
23. Monteiro, N.R.; Ribeiro, B.; Arrais, J.P. Deep Neural Network Architecture for Drug-Target Interaction Prediction. In Proceedings of the International Conference on Artificial Neural Networks, Munich, Germany, 17–19 September 2019; Springer: Cham, Switzerland, 2019.
24. Khan, S.U.; Haq, I.U.; Rho, S.; Baik, S.W.; Lee, M.Y. Cover the Violence: A Novel Deep-Learning-Based Approach towards Violence-Detection in Movies. *Appl. Sci.* **2019**, *9*, 4963. [CrossRef]
25. Haq, I.U.; Muhammad, K.; Ullah, A.; Baik, S.W. DeepStar: Detecting starring characters in movies. *IEEE Access* **2019**, *7*, 9265–9272. [CrossRef]
26. Ullah, A.; Muhammad, K.; Del Ser, J.; Baik, S.W.; de Albuquerque, V.H.C. Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Trans. Ind. Electron.* **2018**, *66*, 9692–9702. [CrossRef]
27. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
28. Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*; IEEE Press: Linz, Austria, 2001.
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

30. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
31. Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2020**, *20*, 183.
32. Hussain, T.; Muhammad, K.; Ullah, A.; Cao, Z.; Baik, S.W.; de Albuquerque, V.H.C. Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM. *IEEE Trans. Ind. Inform.* **2019**, *16*, 77–86. [CrossRef]
33. Ullah, F.U.M.; Ullah, A.; Haq, I.U.; Rho, S.; Baik, S.W. Short-Term Prediction of Residential Power Energy Consumption via CNN and Multilayer Bi-directional LSTM Networks. *IEEE Access* **2019**. [CrossRef]
34. Tan, F.; Feng, X.; Fang, Z.; Li, M.; Guo, Y.; Jiang, L. Prediction of mitochondrial proteins based on genetic algorithm–partial least squares and support vector machine. *Amino Acids* **2007**, *33*, 669–675. [CrossRef]
35. Bhasin, M.; Raghava, G. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* **2004**, *32* (suppl. 2), W414–W419. [CrossRef]
36. Guda, C.; Fahy, E.; Subramaniam, S. MITOPRED: A genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **2004**, *20*, 1785–1794. [CrossRef] [PubMed]
37. Jiang, L.; Li, M.; Wen, Z.; Wang, K.; Diao, Y. Prediction of mitochondrial proteins using discrete wavelet transform. *Protein J.* **2006**, *25*, 241–249. [CrossRef] [PubMed]