Journal of Bioinformatics and Computational Biology
Vol. 17, No. 3 (2019) 1950019 (18 pages)

DOI: 10.1142/S0219720019500197

PyPredT6: A python-based prediction tool for identification of Type VI effector proteins

Rishika Sen*, Losiana Nayak† and Rajat K. De‡

Machine Intelligence Unit
Indian Statistical Institute, 103 B.T. Road
Kolkata-700108, India
**rishikasen_r@isical.ac.in*
†*losiana_t@isical.ac.in*
‡*rajat@isical.ac.in*

Received 5 November 2018
Accepted 25 March 2019
Published 24 June 2019

Prediction of effector proteins is of paramount importance due to their crucial role as first-line invaders while establishing a pathogen-host interaction, often leading to infection of the host. Prediction of T6 effector proteins is a new challenge since the discovery of T6 Secretion System and the unique nature of the particular secretion system. In this paper, we have first designed a Python-based standalone tool, called PyPredT6, to predict T6 effector proteins. A total of 873 unique features has been extracted from the peptide and nucleotide sequences of the experimentally verified effector proteins. Based on these features and using machine learning algorithms, we have performed *in silico* prediction of T6 effector proteins in *Vibrio cholerae* and *Yersinia pestis* to establish the applicability of PyPredT6. PyPredT6 is available at http://projectphd.droppages.com/PyPredT6.html.

Keywords: Host-pathogen interactions; Type VI secretion system (T6SS); gram-negative bacteria; multilayer perceptron; support vector machine; pathogen informatics.

1. Introduction

Gram-negative bacteria have six different secretion systems, viz, Type I (T1SS), Type II (T2SS), Type III (T3SS), Type IV (T4SS), Type V (T5SS) and Type VI (T6SS) secretion systems.[1] These systems facilitate the transfer of certain proteins, known as "effector proteins", required for bacterial growth and infection in the host environment. The component proteins of secretion systems, the transferred proteins and other bacterial protein components, like the flagellar proteins which help in affecting bacterial infection in the host, are known as "effector proteins". The effector proteins play important role in bacterial pathogenesis[2] and competition in host.[3] Among the secretion systems, T6SS had been discovered in 2006. T6SS associated effector proteins in many gram-negative bacteria are yet to be

discovered. *In silico* prediction of these proteins will facilitate faster experimental validation and provide clear information regarding pathogen invasion mechanisms via T6SS.

T6SS is a phage-tail-spike-like injectisome. The injectisome releases effector proteins directly into the cytoplasm of host cells.[4] Genes for T6SS components have been found in proteobacteria, planctomycetes, and acidobacteria. T6SS has also been found in pathogenic species, viz., *V. cholerae*, *E. tarda*, *P. aeruginosa*, *B. mallei* and *F. tularensis* among others. T6SS has been identified to play a major role in the pathogenesis of *A. hydrophila*.[5] T6SS locus (YPO0499-YPO0516) has been found to play a crucial role in phagocytosis-promoting activity.[6] T6SS has also been discovered in plant pathogens, viz., *A. tumefaciens*, *P. atrosepticum* and *X. oryzae* among others. Genes encoding T6SS have also been found in some non-symbionts such as *M. xanthus*, *D. aromatica* and *R. baltica*, where they may contribute to biofilm formation.

A few attempts have already been made toward *in silico* prediction of effector proteins of T3SS and T4SS.[7] Computational prediction of T3 secreted effector proteins using machine learning techniques has been done previously.[8–13] Prediction of secretion signals in genomes of gram-negative bacteria has been done by Löwer *et al.*[8] The authors have used Support Vector Machine (SVM) and an Artificial Neural Network (ANN) with gradient descent back-propagation learning with momentum and an adaptive learning rate to classify proteins as T3 effector proteins and non-effector proteins. Samudrala *et al.*[9] have predicted using SVM, the mechanism of secreted substrates, and identified conserved secretion signal for T3 secretion systems. SVM has also been applied to N-terminal of amino acid sequences to predict novel T3 effector proteins.[10] Similarly, T3 secreted proteins have been predicted based on the amino acid sequences by Arnold *et al.*[11] The authors have compared the performances of prediction made by naive Bayes (NB) classifier, 1-nearest neighbor, logistic regression, NB multinomial, SVM and voted perceptron methods. Wang *et al.*[12] have predicted T3 effector proteins, using a two-layered ensemble predictor Bastion3, based on the features obtained from N-terminal of the proteins. Xue *et al.*[13] have used deep learning framework to predict T3 effector proteins taking only the first 100 residues for prediction.

Identification of T4 effector proteins has been done on the basis of amino acid composition by Zou *et al.*[14] The authors have used SVM to predict T4 effector proteins with an accuracy of 95.9%. The investigation has separately identified T4A and T4B effector proteins. Identification of T4 effector proteins in *Legionella pneumophila* has been done by using a machine learning approach.[15] The ORFs of the proteins in *Legionella pneumophila* have been classified as either effector or non-effector proteins. Genomic, evolutionary, regulatory networks and pathogenic attributes have been extracted from ORFs so as to identify T4 effector proteins. Xiong *et al.*[16] and Wang *et al.*[17] have predicted T4 effectors using ensemble classifiers based on only C-terminal features. The latter group has developed Bastion4 to predict T4 effectors. McDermott *et al.*[18] summarizes the computational prediction of

T3 and T4 effector proteins, concluding that T3 secretion signals are similar across many different bacteria.

Bastion6, an SVM-based protein predictor, is currently the only available tool for prediction of T6 effector proteins.[19] However, multiple limitations have been noticed in the implementation of Bastion6 in terms of its dataset size, choice of non-effector proteins, choice of the classifier, speed of execution, reliability of the results, functionality of the server, its predicted effectors among others. Apart from Bastion6, Zalguizuri *et al.*[20] and An *et al.*[21] have made an attempt to predict T6 effectors. However, the results were unsatisfactory and they expressed a dire need for specialized models for T6 effector prediction. Moreover, these two investigations have considered the non-effector sets for T3 and T4 together as the non-effector set for T6. As mentioned before, due to multifunctional nature of the proteins, proteins being T3/T4 non-effectors need not necessarily be T6 non-effectors.

Some notable demerits in all of the above investigations are that hypothetical proteins have been used in training data. In some of the investigations, non-effector dataset has been derived by choosing secreted proteins obtained from any of T1SS through T8SS in Gram-negative bacteria. For creating the non-effector set in T3/T4 prediction, secreted proteins of types T1SS through T8SS, except T3/T4, have been taken into consideration. It may result in a non-effector list containing effector proteins since proteins are multi-functional in nature. Moreover, none of the aforesaid investigations have applied feature selection on their datasets, which elevates the risk of over-fitting.

In order to overcome the shortcomings of Bastion6, we have developed PyPredT6, a python-based standalone tool for predicting probable T6 effector proteins using a set of unique 837 features derived from existing biological knowledge, viz., SecReT6[22] and the SecretEPDB[23] databases. These 837 features have been extracted from the nucleotide and the peptide sequences of the experimentally verified sequences obtained from the aforesaid databases. Moreover, we have applied PyPredT6 to the sequences of *Vibrio cholerae* and *Yersinia pestis* to predict probable T6 effector proteins, which have not yet been discovered in both the species. Such a measure can exponentially accelerate the host-pathogen interaction studies of the pathogen informatics research field.

The paper is organized as follows. In Sec. 2, we describe the training data and the feature set in detail along with the technical details of PyPredT6. This is followed by Sec. 3 with the predictions done in *V. cholerae* and *Y. pestis*. A detailed comparison of PyPredT6 with Bastion6 is carried out in Sec. 4. Finally, in Sec. 5, we discuss the applicability and future scopes of PyPredT6 along with some concluding remarks.

2. Methodology

PyPredT6 is a standalone tool that runs on Python 3.6 and above. It has been written using Win Python in a 64-bit machine. PyPredT6 can read nucleotide and amino acid sequences from text files in FASTA format. Prediction is done by

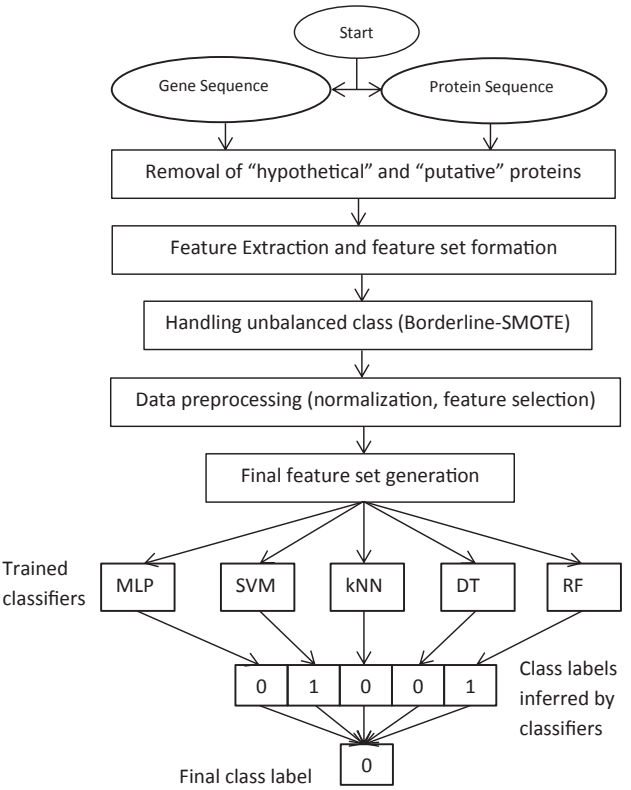


Fig. 1. Methodology for prediction of putative T6 effector proteins. A value of 1 indicates that a protein is pathogenic, while 0 stands for a protein being non-pathogenic. Here an example of final class label of 0 is provided based on majority voting of outcomes of the classifiers.

consensus of Multi-Layer Perceptron (MLP), SVM, k Nearest Neighbor (k-NN), NB and Random Forest (RF) classifiers via voting.[24] A consensus of classifiers is used to build a robust system for the prediction of T6 effector proteins. PyPredT6 code along with the instructions for executing the code is available at http://projectphd.drop-pages.com/PyPredT6.html. The overall methodology followed in designing PyPredT6 is presented as a flowchart in Fig. 1.

2.1. *Training data*

We have accumulated a set of experimentally verified amino acid sequences of T6 effector proteins in different organisms from two databases, viz., SecReT6[a,22] and SecretEPDB[b].[23] The corresponding nucleotide sequences have also been considered. A total of 175 unique effector proteins has been obtained from the databases. The non-effector set has been constructed from the entire genome of non-pathogenic

[a] http://db-mml.sjtu.edu.cn/SecReT6/.

[b] http://secretepdb.erc.monash.edu/.

gram-negative bacteria *Bacteriodes vulgatus*.[25] An argument may arise here that housekeeping proteins of the same pathogenic species (from which effector proteins have been taken) provide a better option for the non-effector proteins. However, in prokaryotes, genes are often found to be multi-functional in nature.[26,27] In order to avoid a housekeeping gene of the same species, which has some kind of direct or indirect association with an effector gene,[28] here we have considered a different non-pathogenic gram-negative bacteria (*B. vulgatus*) which lives in human gut. It has to be mentioned here that the T6 effector protein set considered here comprises proteins from multiple gram-negative species.

A set of 4183 genes and their corresponding proteins of *B. vulgatus* has been obtained from KEGG.[c] Proteins annotated as "putative","hypothetical" and "uncharacterized" have been removed from the set as no physical, genetic or functional annotation is available for such proteins. Thus a total of 1063 putative, 1572 hypothetical and 51 uncharacterized proteins have been removed from the initial set. Finally, we have considered 1497 non-effector proteins of *B. vulgatus*.

2.2. *Feature set*

In this section, we derive nucleotide and amino acid-based features from the sequences of T6 effector and non-effector proteins. A schematic representation of the feature set has been given in Fig. 2.

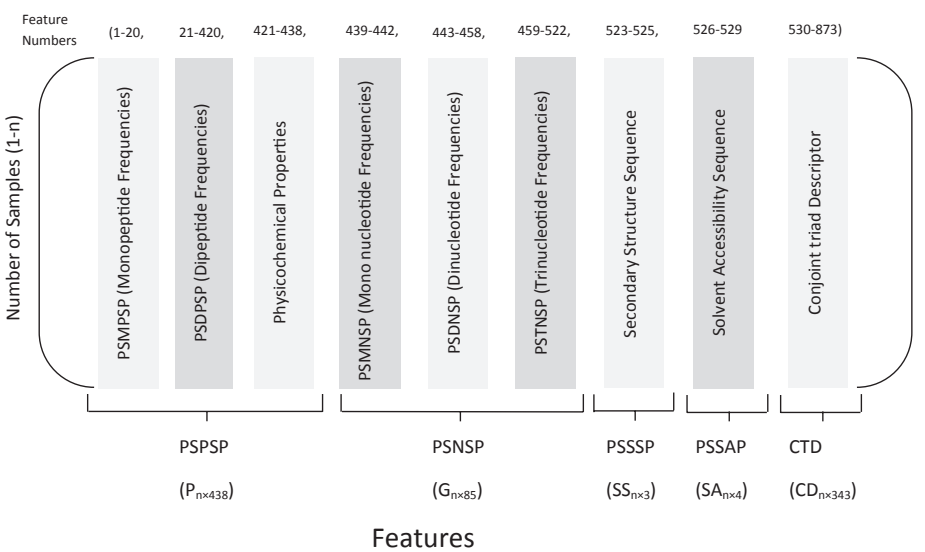


Fig. 2. The structure of the feature matrix where $G_{n\times 85}$: gene feature matrix, $P_{n\times 438}$: protein feature matrix, $\mathrm{CTD}_{n\times 343}$: conjoint triad descriptor matrix, $SS_{n\times 3}$: secondary structure feature matrix, and $SA_{n\times 4}$: solvent accessibility feature matrix.

[c] https://www.genome.jp/dbget-bin/get_linkdb?-t+genes+gn:T00546.

Position specific nucleotide sequence profile (PSNSP): These features have been extracted from the nucleotide sequences of the genes. The percentage composition of 4 mono-nucleotides (A, T, G, C) in a gene, i.e. the percentage of each of A, T, G and C with respect to the total number of nucleotides in the sequence of a gene form position-specific mononucleotide sequence profile (PSMNSP). Likewise, the percentage composition of 16 di-nucleotides (AA, AT, AG, ..., and others) with respect to the total number of dinucleotides in the gene sequence form the position-specific dinucleotide sequence profile (PSDNSP). The percentage composition of 64 tri-nucleotides (AAA, AAT, AAG, ..., and others) with respect to the total number of triplets form position-specific trinucleotide sequence profiles (PSTNSP). Thus, PSNSP of a gene comprises PSMNSP (4 features), PSDNSP (16 features), PSTNSP (64 features), and GC content. In this way, we have got 85 features for a gene. These features altogether constitute the gene feature matrix $G_{n\times 85}$ for n gene sequences.

Position specific peptide sequence profile (PSPSP): These features have been extracted from the protein sequences. The percentage composition of 20 single amino acids (A, G, H, ..., and others) in a protein, i.e. the percentage of each of A, G, E, V, I, L, F, P, Y, M, T, S, H, N, Q, W, R, K, D and C with respect to the total number of peptides in the sequence of the protein form the PSMPSP. Likewise, the percentage composition of 400 di-peptides (AA, AG, AH, ..., and others) with respect to the total number of dipeptides in the protein sequence form the position-specific dipeptide sequence profile (PSDPSP). PSPSP comprises PSMPSP (20 features), PSDPSP (400 features) and 18 physicochemical properties. The different classes of amino acids corresponding to the physicochemical properties considered are charged (D, E, K, H, and R), aliphatic (I, L, and V), aromatic (F, H, W, and Y), polar (D, E, R, K, Q, and N), neutral (A, G, H, P, S, T, and Y), hydrophobic (C, F, I, L, M, V, and W), positively charged (K, R, and H), negatively charged (D and E), tiny (A, C, D, G, S, and T), small (E, H, I, L, K, M, N, P, Q, and V), large (F, R, W, and Y), transmembrane amino acid (I, L, V, A), dipole < 1.0 (A, G, V, I, L, F, P), $1.0 <$ dipole < 2.0 (Y, M, T, S), $2.0 <$ dipole < 3.0 (H, N, Q, W), dipole > 3.0 (R, K), and dipole > 3.0 with opposite orientation (D, E, C).[29,30] In order to calculate the physicochemical properties, the sum of the percentage composition of the amino acids belonging to each of these 18 classes is considered. These features altogether form the protein feature matrix '$P_{n\times 438}$' for n protein sequences.

Position specific secondary structure profile (PSSSP): We have considered three types of secondary structures of a protein, i.e. helix (H), coil (C) and sheets (E) to form the matrix '$SS_{n\times 3}$' for n protein sequences. The amino acids E, A, L, M, Q, K, R and H form helix in secondary structure format. Likewise, the amino acids G, N, P, S and D are known to form coil. Lastly, the amino acids V, I, Y, C, W, F and T collectively form sheet. In order to find the secondary structure composition, the sum of the percentage composition of the amino acids belonging to helix, coil and sheets are considered.

Presence of helices or coiled coils or sheets as domains in effector proteins has its own significance. Helices confer evolvability,[31] attachment to host membrane,[32] actin nucleation.[33] Crystal structures of the effector domains from two oomycete RXLR proteins, *Phytophthora capsici* AVR3a11 and *Phytophthora infestans* PexRD2 reveal a conserved core α-helical fold.[31] The fold exists in $\sim 44\%$ of the annotated *Phytophthora* RXLR effectors, both as a single domain and in tandem repeats of up to 11 units.[31] According to Boutemy *et al.*, the core α-helical fold displays the evolution of effector proteins to gain new virulence functions and/or evade the host immune system by insertion/deletions in loop regions between α-helices, extensions to the N and C termini, amino acid replacements in surface residues, tandem domain duplication, and oligomerization. A study by Weigele *et al.*[32] suggested that *Shigella* IpgB1 utilizes an amphipathic helix enriched with basic residues to interact directly with acidic phospholipids of host cell membrane. *Vibrio* T3 effector protein VopL contains three closely spaced WH2 domains (short 17-22 residues regions nearly always found in tandem and forming an N-terminal helix with a conserved downstream LKKV motif) which take part in actin stress fibre formation by directly nucleating actin filaments.[33]

Coiled coil (alpha-helices coiled together) domains impart membrane attachment[34] and immunity.[35] Knodler *et al.*[34] suggested that coiled-coil domains are prevalent in virulence-associated proteins, including T3 effectors in *Salmonella enterica serovar Typhimurium*. These domains may represent a common membrane-targeting determinant for *Salmonella* T3 effectors.[34] Distinct regions of the *Pseudomonas syringae* coiled-coil effector AvrRps4 are required for activation of immunity.[35] Presence of β-sheets in effector proteins facilitate host-pathogen interaction.[36] *Salmonella* effector Protein SopB forms an inter-molecular-sheet with Cdc42 of the host organism.[36]

Position specific solvent accessibility profile (PSSAP): The solvent accessibility feature of an amino acid can be very buried (B), somewhat buried (b), very exposed (E), and somewhat exposed (e). We have considered these four features to form the solvent accessibility feature matrix '$SA_{n\times 4}$' for n protein sequences. The solvent accessibility has been calculated using the DSSP[37] program. An amino acid is said to be very buried (B) when its accessibility is at most 4%, somewhat buried (b) when accessibility is between 4% and 25%, somewhat exposed (e) when accessibility is between 25% and 50% and very exposed (E) when accessibility is more than 50%.[38,39] Amino acids that can be characterized as very buried are A, L, F, C, I, V; somewhat buried amino acids are W, M, S, P, T, H and Y. Similarly, amino acids that are exposed are Q, E, D; and amino acids that are somewhat exposed are R, K, N and G. In order to calculate the solvent accessibility profile, the sum of the percentage composition of the amino acids being very buried, somewhat buried, very exposed and somewhat exposed are considered.

The solvent accessibility of a protein has an influence on their structure which in turn influences their functionality.[40] The extent to which the structure of proteins has an impact on their function is shown by the effect of changes in the structure of a

protein. Any change to a protein at any structural level, including slight changes in the folding and shape of the protein, may render it non-functional.[41] The solvent accessibility feature of proteins is often used for identifying gram negative effector proteins.[10,42]

Conjoint Triad Descriptors (CTD): These features have been extracted from the amino acid sequences. The conjoint triad descriptors consider a group of three consecutive amino acids (triads) with respect to the protein sequences and their assigned groups depending on the classification based on dipole scale of each amino acid and volumes of side chains.[43] The distribution of the amino acids in each group has been given in Table 1. There are seven classes into which 20 amino acids can be placed. We have considered three consecutive amino acids (triplet) for further calculation. Considering three consecutive amino acids, each of the three amino acids will belong to one of the groups. The combination of the groups for three consecutive amino acids looks like $[3, 1, 7]$, for example, if these three amino acids are in Groups $3, 1$ and 7 respectively. Since three positions have been taken into consideration and each amino acid can belong to a single group, there can be one of 343 $(= 7 \times 7 \times 7)$ possible groups for each triplet of amino acids. The frequency of triplets belonging to each of these 343 combinations of groups are taken into account to obtain the final matrix of order $n \times 343$ (CTD), where n is the total number of sequences. The frequency of each triad belonging to one of the combinations of groups forms the CTD. For example, considering the peptide sequence *IMFTLED*. The combinations of *IMF*, *MFT* and *FTL* are $[2, 3, 2]$, $[3, 2, 3]$ and $[2, 3, 2]$. Hence, the frequencies of $[2, 3, 2]$, $[3, 2, 3]$, $[3, 2, 6]$ and $[2, 6, 6]$ are 2, 1, 1 and 1 respectively, while the rest of the groups have frequencies of 0.

The features $G_{n\times 85}$, $P_{n\times 438}$, $SS_{n\times 3}$, $SA_{n\times 4}$ and $\text{CTD}_{n\times 343}$ have been combined to form a single feature matrix $F_{n\times(85+438+3+4+343=873)}$ as shown in Fig. 2 for 873 features corresponding to each of the n genes/proteins. We have also generated some other features but could not consider them due to their non-conclusiveness. We have not included information regarding Pfam domains, palindrome sequences, nucleotide analysis of N- and C-terminals in our analysis due to their insignificant contribution[19] in distinguishing between effectors and non-effectors. We could not find any universal or major represented Pfam domain for the dataset taken from SecRet6 and

Table 1. Summary of the distribution of amino acids based on their dipole and volumes of the side chains.

Group	Amino acid
1	Alanine (A), Glycine (G), Valine (V)
2	Isoleucine (I), Leucine (L), Phenylalanine (F), Proline (P)
3	Tyrosine (Y), Methionine (M), Threonine (T), Threonine (S)
4	Histidine (H), Asparagine (N), Glutamine (Q), Tryptophan (W)
5	Arginine (R), Lysine (K)
6	Aspartic acid (D), Glutamic acid (E)
7	Cysteine (C)

SecretEPDB databases. We have not found any common palindrome sequence in the candidate genes. Moreover, we have considered amino acid composition, dipeptide composition and physicochemical properties of N- and C-terminals of the amino acid sequences, but could not find any significant differentiating factor between the effector and the non-effector proteins.

2.3. *Secondary structure-based feature analysis of the effectors and non-effectors*

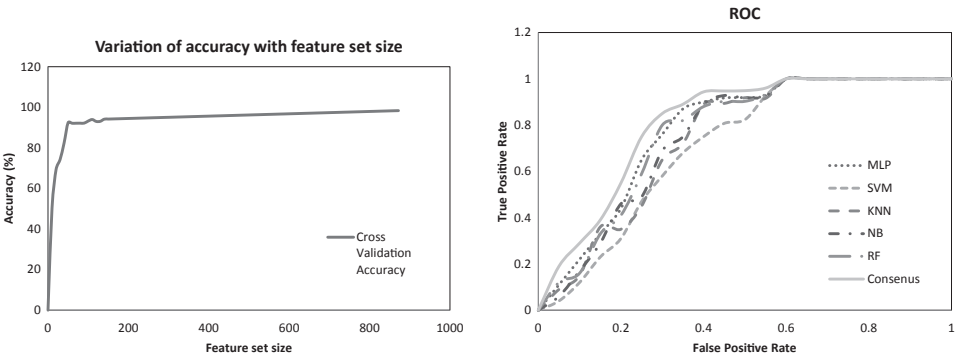
The secondary structure composition of the effector and non-effector proteins displays contrast in distribution. A considerable difference has been noticed in the distribution of α-helices and β-sheets in both the categories of proteins. As given in Table 2, the overall percentage of helices in effector proteins is less than that in non-effector proteins. Similarly, the percentage of β sheets is more in effector proteins than in non-effector proteins. Statistical analysis of the correlation between α-helices and β-sheets in effectors shows a strong positive correlation among them with a p-value < 0.05 and a Pearson correlation coefficient $r = 0.88$. Such a correlational significance is absent in non-effector proteins. Although we could not establish any immediate relevance of such a finding, the stark contrast in the distribution pattern of α-helices and β-sheets needs further investigation.

2.4. *Preprocessing of feature sets*

The cardinalities of the sets of effector and non-effector proteins are unbalanced, i.e. the number of samples in the effector class is considerably less than the number of samples in the non-effector class due to the unavailability of more experimentally verified T6 effector proteins. An equal sized sets of effector and non-effector proteins need to be considered to avoid unequal class distribution and a biased classifier.[44] In order to do so, we have over-sampled the dataset using Borderline-SMOTE over-sampling method[45] so that cardinality of the minority class (T6 effector proteins) has become approximately equal to that of the majority class (non-effector proteins). Borderline-SMOTE over-samples only the borderline examples of the minority class. For every minority sample, its K-NN of the same class have been found, followed by the selection of some random samples from them according to the over-sampling rate. Hence, the new synthetic examples are generated along the boundary of the minority class and its selected nearest neighbors. This is followed by standardization of the features by subtracting them from the mean followed by scaling them to unit variance.

Table 2. Composition of secondary structures in the experimentally verified T6 effector proteins.

Class	Coil (in %)	Helix (in %)	Sheet (in %)
Effector proteins	45.48	19.69	33.42
Non-Effector proteins	39.11	40.98	18.43



(a) Accuracy versus Feature set size

(b) ROC Curve

Fig. 3. Performance of PyPredT6. (a)-represents the variation of accuracy with the feature set size. (b)-represents the Receiver Operating Characteristic curve comparing the individual performances of the five classifiers and the consensus of classifiers. As visible, consensus of the five classifiers gives a better prediction result compared to the individual classifiers.

We have used Gini impurity index in a randomized decision tree[46] for feature selection on various sub-samples of the dataset to avoid over-fitting and improve the predictive accuracy of the classifiers. The classifiers have been tested on 10, 20, 30, . . . ,850, 860, 873 features with 10 most significant features getting added in each iteration.[47] The performance of the predictors has been recorded for such datasets of different feature size and plotted in Fig. 3(a). It has been found that out of 873 features, 51 most significant features with respect to Gini impurity index are of high importance. The classifier has been seen to have relatively stable with negligible difference in accuracy. The size of the feature set has been increased from 51 most significant features achieving an accuracy of 92.13%, to include all the 873 features resulting in an accuracy of 95.36%. The variation of performances of the classifiers on dataset of different sizes has been depicted in Fig. 3(a). To avoid over-fitting, only these highly important features have been used for further classification and prediction.

2.5. *Performance assessment of PyPredT6*

PyPredT6 uses the consensus of MLP, SVM, k-NN, NB and RF classifiers. It decides whether an unknown protein is a T6 effector or not using the method of majority voting. In this respect, the predicted values of all the five classifiers have been taken into consideration. The class predicted by majority of classifiers has been considered as the final class for a certain protein.

The MLP classifier has six hidden layers having the activation function ReLU for each node. The output layer nodes have the sigmoid activation function. The SVM classifier uses RBF kernel. The k-NN classifier has considered $k = 10$. The performance of the different classifiers has been assessed by Accuracy, Sensitivity,

Table 3. Summary of performance (in %-age) of the five classifiers with 10-fold cross-validation. The tabulated values are the 50-fold average for each of the classifiers.

Classifier	Accuracy	Sensitivity	Specificity	Fscore	Gmean
Multilayer perceptron	87.52	88.35	86.15	84.03	85.35
SVM	84.57	80.80	87.70	81.54	87.24
K-NN	88.05	81.82	87.25	82.56	84.69
NB	89.42	81.27	88.45	85.42	84.63
RF	76.15	81.25	83.75	86.32	83.45
Consensus	**92.15**	**91.25**	**90.75**	**87.45**	**88.39**

Specificity, F Score and G-Mean, which are defined in Sec. 1 of Supplementary Information.

The individual performance and the consensus performance of the classifiers have been tabulated in Table 3. The ROC curve[48] for the same has been depicted in Fig. 3 (b). As evident from the table and the plot, the consensus of the classifiers gives a better performance in identifying an unknown protein to be a T6 effector or a non-effector.

3. Application of PyPredT6 on *Vibrio Cholerae* and *Yersinia pestis* Proteins

The consensus of the five classifiers has been used to predict probable T6 effector proteins in *V. cholerae* and *Y. pestis*. The amino acid sequences for both the species have been obtained from Biocyc.[49] We have collected 2736 nucleotide and their respective amino acid sequences of *V. cholerae*. We have also collected 3850 nucleotide and their respective amino acid sequences of *Y. pestis*.

Out of 2736 proteins of *V. cholerae* (Chromosome 1: Strain O1 biovar El Tor str. N16961,[d] version 21.1), 30 proteins have been selected by PyPredT6 to be probable effectors. For *Y. pestis* (Strain Pestoides F, version 21.1[e]), out of 3850 proteins, 42 proteins have been selected to be effectors. Here, the predicted probable T6 effector proteins of our two test species have been discussed after secondary structure based feature filtering. In order to biologically validate these proteins, we have considered their gene ontology information (as listed in UniProtKB[f,50]) and information from the existing literature. In this way, we have established a direct/indirect relation with virulence and pathogenesis of a few of these proteins. The literature-based validation of probable T6 effector proteins predicted by PyPredT6 has been furnished in Secs. 3 and 4 of Supplementary Information, respectively. A tabulated form of the same has been furnished in http://projectphd.droppages.com/PyPredT6.html.

[d] https://biocyc.org/VCHO/organism-summary?object=VCHO.

[e] https://biocyc.org/YPES386656/organism-summary?object=YPES386656.

[f] http://www.uniprot.org/.

4. Comparison of PyPredT6 with Bastion6

Bastion6[19] is the only other tool that attempts to predict T6 effector proteins. However, multiple limitations have been observed in the tool. Bastion6 has extracted experimentally verified data from SecretEPDB while PyPredT6 has taken into consideration data from both SecReT6 and SecretEPDB. The training dataset of Bastion6 is imbalanced (consisting of 20 effector proteins and 200 non-effector protein samples). A low number of positive samples and a high number of features for prediction indicate Bastion6 as a probable over-fitted classifier. PyPredT6, on the other hand, has a positive set of 175 effectors and 1497 non-effectors. In order to do away with the problem of an imbalanced dataset, oversampling has been performed using borderline-SMOTE technique.

The non-effector samples (negative dataset) of Bastion6 have been taken from two sources - the non-effector proteins of Zou *et al.*[14] and those in *Vibrio parahaemolyticus*. The non-effector proteins from Zou *et al.* comprise those which are not T4 effectors. Due to multi-functional nature of prokaryotic genes,[27] this may not be a safe approach. Proteins which are not T4 effectors may have an association with T6SS machinery. On the other hand, *Vibrio parahaemolyticus* is a pathogenic gram-negative bacteria.[51] Hence, there arises a risk of considering an effector protein as a non-effector in the negative dataset. In order to avoid all these issues, PyPredT6 has taken the non-effector dataset, from an experimentally verified non-pathogenic organism.

Bastion6 has considered 1096 features in total with a sample size of 220. Given the high number of features and the limited number of samples, the dire need for feature selection and oversampling is noticed, which if not done, will lead to over-fitting.[52] The avoidance of using a feature selection and an oversampling method for Bastion6 indicates over-fitting.[53] Hence, PyPredT6 has incorporated feature selection along with an oversampled large training set with the intention to avoid over-fitting. PyPredT6 has used the most significant 51 features on an oversampled dataset of size 2994. For Bastion6, the values of accuracy (94.3%), sensitivity (100%) and specificity (88%) have a considerable variance among the measurements, indicating quite an unstable performance. PyPredT6, however, displays a stable performance over accuracy (89.15%), sensitivity (91.25%) and specificity (90.75%), with a considerably low variance. Such high variance in the above measurements for Bastion6 may indicate overfitting. Besides, Bastion6 has displayed a low specificity indicating a high number of false positives and low number of true negatives.

Bastion6 has considered the result of a single SVM classifier to predict the effector proteins, whereas PyPredT6 takes into account the prediction of five classifiers, and uses a voting method to obtain the final class label of the sample protein. An extensive study has been performed to measure the CPU time of PyPredT6. The summary of the study has been given in Table S1 in Supplementary Information. CPU time for PyPredT6 on three random datasets containing 10, 20 and 30 sequences have been recorded. Here a single sequence refers to a pair of nucleotide

and the corresponding amino acid sequences. For each set of sequences, the time needed for training PyPredT6 (T_T) with the feature set of experimentally verified effectors and the time required to extract features from unknown sequences (T_E) have been recorded. The average of total execution time ($T_S = T_E + T_T$) of PyPredT6 (5.24 min on a 32GB RAM, 64 bit Windows operating system) on the aforesaid three datasets is considerably less than that of Bastion6 (29.6 min on Bastion6 server). As observed from the table, the training time is nearly constant with an average of 314.61 s, while the average time for feature extraction for a single sequence is approximately 0.0751 s.

PyPredT6 is a standalone application, which can be downloaded from the website (http://projectphd.droppages.com/PyPredT6.html/). Bastion6 is restricted to process less than 500 sequences per job with amino acid count between 50 and 5000. PyPredT6, on the other hand, does not have any limit on the length or the number of sequences.

From a total of 6586 proteins of two species, we have predicted 72 effector proteins. PyPredT6 aims to reduce the true negatives while predicting the effector proteins. Bastion6 has considered 12 species for predicting the effector proteins.[g] Among them, it has validated two proteins as probable effectors, while we have validated the possibility of all the predicted 72 proteins for being probable effector proteins. A summary of the comparison has been given in Table 4.

In order to assess and benchmark the performance of PyPredT6 and Bastion6, we have created three sets Sets 1, 2 and 3 (Supplementary Information), of independent

Table 4. Summary of the fundamental differences between PyPredT6 and Bastion6.

Field	PyPredT6	Bastion6
Database	SecRet6 and SecretEPDB	SecretEPDB
Sample size	175 effectors, 1497 non-effectors	20 effectors, 200 non-effectors
Non-effector set	Entire genome of non-pathogenic gram-negative bacteria *Bacteriodes vulgatus*[25]	Proteins which are non-effectors with respect to T4 effectors, and from *Vibrio parahaemolyticus*, a pathogenic gram-negative bacteria
Feature types	Peptide and nucleotide features	Peptide features
Oversampling technique	borderline-SMOTE	None applied
Feature selection technique	Randomized decision tree using Gini impurity index	None applied, indicating over-fitting
Classifier	Consensus of MLP, SVM, KNN, NB, RF	SVM
Execution time	5.24 minutes	29.60 minutes
Input constraints	Able to handle any number of sequences of any size	Unable to handle more than 500 sequences per job, length of each sequence between 50 and 5000.
Performance comparison	89.15% (Acc), 91.25% (Sen), 90.75% (Spe)	94.3% (Acc), 100% (Sen), 88% (Spe)

[g]List available in the Supplementary Information of the original paper.

non-overlapping effectors and non-effectors data extracted from the public databases and the literature, for comparing the predictive power of PyPredT6 and Bastion6. The first dataset, Set 1 (Table S2), has been constructed taking all the T6 effector proteins of *Edwardsiella tarda* from Genbank. The second set, Set 2 (Table S3), has been constructed using a handful of proteins of *Homo sapiens* obtained from Genbank, which cannot be effectors. The third dataset, Set 3 (Table S4), consists of T6 effector proteins accumulated from the literature. PyPredT6 has shown promising results while predicting effector and non-effector proteins from a pool of unknown proteins. Bastion6, on the other hand, has been unable to provide any conclusive result on classification of these proteins belonging to Sets 1 and 2. For Set 3, Bastion6 has been able to predict 6 out of 10 proteins correctly, while PyPredT6 has been able to predict 9 out of 10 proteins correctly. The Tables S2, S3 and S4 are available in the Supplementary Information.

5. Conclusions

Prediction of effector proteins from bacterial genome information is important for the analysis of their secretion systems' role in pathogenesis. Here we have developed a standalone python tool, called PyPredT6, for prediction of probable T6 effector proteins based on consensus of five classifiers. PyPredT6 extracts a feature set having 873 features from nucleotide and amino acid sequences of experimentally verified T6 effector proteins. PyPredT6 has predicted 42 proteins out of 3850 proteins from *Y. pestis* and 30 proteins out of 2736 proteins from *V. cholerae* as effectors. We have analyzed these proteins for being putative T6 effector proteins in a limited capacity. PyPredT6 offers users to check whether a protein is a T6 effector or not. A more detailed biological validation for each putative candidate gene is essential, which forms a scope for further study. The methodology can be extended to other pathogens, whose genomes and proteomes are either partially or fully mapped.

Acknowledgments

LN acknowledges the University Grants Commission, India for providing her a UGC Post-Doctoral Fellowship (No. F.15-1/2013-14/PDFWM-2013-14-GE-ORI-19068-(SA-II)). RS has conceptualized, carried out the experimental work and has prepared the first draft of the paper. LN has helped in biological validation of the probable predicted T6 effector proteins. RKD has given technical inputs and thoroughly corrected the paper.

References

1. Costa TR, Felisberto-Rodrigues C, Meir A, Prevost MS, Redzej A, Trokter M, Waksman G, Secretion systems in Gram-negative bacteria: Structural and mechanistic insights, *Nat Rev Microbiol* **13**:343–359, 2015.
2. Tillotson GS, Tillotson J, Bacterial secreted proteins: Secretory mechanisms and role in pathogenesis, *Expert Rev Anti-infective Ther* **7**:691–693, 2009.

3. Sana TG, Lugo KA, Monack DM, T6SS: The bacterial "fight club" in the host gut, *PLoS Pathogens* **13**:e1006325, 2017.
4. Cascales E, The Type VI secretion toolkit, *EMBO Rep* **9**:735–741, 2008.
5. Suarez G, Sierra JC, Sha J, Wang S, Erova TE, Fadl AA, Foltz SM, Horneman AJ, Chopra AK, Molecular characterization of a functional type VI secretion system from a clinical isolate of aeromonas hydrophila, *Microbial Pathogenesis* **44**:344–361, 2008.
6. Robinson JB *et al.*, Evaluation of a Yersinia pestis mutant impaired in a thermoregulated type VI-like secretion system in flea, macrophage and murine models, *Microbial Pathogenesis* **47**:243–251, 2009.
7. Sen R, Nayak L, De R, A review on host–pathogen interactions: Classification and prediction, *Eur J Clin Microbiol Infectious Diseases* 1–19, 2016.
8. Löwer M, Schneider G, Prediction of Type III secretion signals in genomes of gram-negative bacteria, *PloS One* **4**:e5917, 2009.
9. Samudrala R, Heffron F, McDermott JE, Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for Type III secretion systems, *PLoS Pathogens* **5**:e1000375, 2009.
10. Yang Y, Zhao J, Morgan RL, Ma W, Jiang T, Computational prediction of Type III secreted proteins from gram-negative bacteria, *BMC Bioinf* **11**:1, 2010.
11. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes H-W, Horn M, Rattei T, Sequence-based prediction of Type III secreted proteins, *PLoS Pathogens* **5**:e1000376, 2009.
12. Wang J *et al.*, Bastion3: A two-layer ensemble predictor of type iii secreted effectors, *Bioinf* 2018.
13. Xue L, Tang B, Chen W, Luo J, Deept3: Deep convolutional neural networks accurately identify gram-negative bacterial type iii secreted effectors using the n-terminal sequence, *Bioinf* 2018.
14. Zou L, Nan C, Hu F, Accurate prediction of bacterial Type IV secreted effectors using amino acid composition and PSSM profiles, *Bioinf* **29**:3135–3142, 2013.
15. Burstein D, Zusman T, Degtyar E, Viner R, Segal G, Pupko T, Genome-scale identification of Legionella pneumophila effectors using a machine learning approach, *PLoS Pathogens* **5**:e1000508, 2009.
16. Xiong Y, Wang Q, Yang J, Zhu X, Wei D, Predt4se-stack: Prediction of bacterial type iv secreted effectors from protein sequences using a stacked ensemble method, *Front Microbiol* **9**:2571, 2018.
17. Wang J *et al.*, Systematic analysis and prediction of type iv secreted effector proteins by machine learning approaches, *Briefings Bioinf* 2017.
18. McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, Sharp D, Adkins JN, Samudrala R, Heffron F, Computational prediction of type iii and iv secreted effectors in gram-negative bacteria, *Infection and Immunity* **79**:23–32, 2011.
19. Wang J *et al.*, Bastion6: A bioinformatics approach for accurate prediction of type VI secreted effectors, *Bioinf* 2018.
20. Zalguizuri A, Caetano-Anollés G, Lepek VC, Phylogenetic profiling, an untapped resource for the prediction of secreted proteins and its complementation with sequence-based classifiers in bacterial type iii, iv and vi secretion systems, *Briefings Bioinf* 2018.
21. An Y, Wang J, Li C, Leier A, Marquez-Lago T, Wilksch J, Zhang Y, Webb GI, Song J, Lithgow T, Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI, *Briefings Bioinf* bbw100, 2016.
22. Li J, Yao Y, Xu HH, Hao L, Deng Z, Rajakumar K, Ou H-Y, SecReT6: A web-based resource for Type VI secretion systems found in bacteria, *Environ Microbiol* 2015.

23. An Y *et al.*, SecretEPDB: A comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems, *Sci Rep* **7**:41031, 2017.
24. Ala'raj M, Abbod MF, Classifiers consensus system approach for credit scoring, *Knowl.-Based Syst* **104**:89–105, 2016.
25. Haller D, Holt L, Kim SC, Schwabe RF, Sartor RB, Jobin C, Transforming growth factor-β1 inhibits non-pathogenic gramnegative bacteria-induced NF-κb recruitment to the interleukin-6 gene promoter in intestinal epithelial cells through modulation of histone acetylation, *J Biol Chem* **278**:23851–23860, 2003.
26. Pereira PR, Fernandes LG, de Souza GO, Vasconcellos SA, Heinemann MB, Romero EC, Nascimento AL, Multifunctional and redundant roles of leptospira interrogans proteins in bacterial-adhesion and fibrin clotting inhibition, *Int J Med Microbiol* **307**:297–310, 2017.
27. Kaimer C, Graumann PL, Players between the worlds: Multifunctional DNA translocases, *Curr Opin Microbiol* **14**:719–725, 2011.
28. Santin YG, Cascales E, Domestication of a housekeeping transglycosylase for assembly of a type VI secretion system, *EMBO Rep* **18**:138–149, 2017.
29. Landolt-Marticorena C, Williams KA, Deber CM, Reithmeier RA, Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins, 1993.
30. Meher PK, Sahu TK, Saini V, Rao AR, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chous general PseAAC, *Sci Rep* **7**, 2017.
31. Boutemy LS, King SR, Win J, Hughes RK, Clarke TA, Blumenschein TM, Kamoun S, Banfield MJ, Structures of Phytophthora RXLR effector proteins a conserved but adaptable fold underpins functional diversity, *J Biol Chem* **286**:35834–35842, 2011.
32. Weigele BA, Orchard RC, Jimenez A, Cox GW, Alto NM, A systematic exploration of the interactions between bacterial effector proteins and host cell membranes, *Nat Commun* **8**:532, 2017.
33. Dean P, Functional domains and motifs of bacterial type III effector proteins and their roles in infection, *FEMS Microbiol Rev* **35**:1100–1125, 2011.
34. Knodler LA, Ibarra JA, Pérez-Rueda E, Yip CK, Steele-Mortimer O, Coiled-coil domains enhance the membrane association of Salmonella type III effectors, *Cellular Microbiol* **13**:1497–1517, 2011.
35. Sohn KH, Hughes RK, Piquerez SJ, Jones JD, Banfield MJ, Distinct regions of the Pseudomonas syringae coiled-coil effector AvrRps4 are required for activation of immunity, *Proc National Academy of Sci* **109**:16371–16376, 2012.
36. Burkinshaw BJ, Prehna G, Worrall LJ, Strynadka NC, Structure of salmonella effector protein SopB N-terminal domain in complex with host Rho GTPase Cdc42, *J Biol Chem* **287**:13348–13355, 2012.
37. Kabsch W, Sander C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**:2577–2637, 1983.
38. Mucchielli-Giorgi M-H, Hazout S, Tuffry P, PredAcc: Prediction of solvent accessibility, *Bioinf* **15**:176–177, 1999.
39. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH, Hydrophobicity of amino acid residues in globular proteins, *Science* **229**:834–838, 1985.
40. Wang Z, Moult J, SNPs, protein structure, and disease, *Human Mutation* **17**:263–270, 2001.
41. Whitford D, *Proteins: Structure and Function*, John Wiley & Sons, 2013.
42. Yang X, Guo Y, Luo J, Pu X, Li M, Effective identification of Gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles, *PloS One* **8**:e84439, 2013.

43. Chen Z *et al.*, iFeature: A python package and web server for features extraction and selection from protein and peptide sequences, *Bioinf* **1**, 2018.
44. Ganganwar V, An overview of classification algorithms for imbalanced datasets, *Int J Emerging Technol Adv Eng* **2**:42–47, 2012.
45. Han H, Wang W-Y, Mao B-H, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, in *Int Conf Intelligent Comput*, Springer, pp. 878–887, 2005.
46. Venables WN, Ripley BD, Tree-based methods, in *Mod Appl Statistics with S*, Springer, pp. 251–269, 2002.
47. Díaz-Uriarte R, De Andres SA, Gene selection and classification of microarray data using random forest, *BMC Bioinf* **7**:3, 2006.
48. Hanley JA, McNeil BJ, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiol* **143**:29–36, 1982.
49. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N, Expansion of the BioCyc collection of pathway-genome databases to 160 genomes, *Nucleic Acids Res* **33**:6083–6089, 2005.
50. Consortium U *et al.*, UniProt: A hub for protein information, *Nucleic Acids Res* gku989, 2014.
51. Letchumanan V, Chan K-G, Lee L-H, Vibrio parahaemolyticus: A review on the pathogenesis, prevalence, and advance molecular identification techniques, *Front Microbiol* **5**:705, 2014.
52. Stockwell DR, Peterson AT, Effects of sample size on accuracy of species distribution models, *Ecological Model* **148**:1–13, 2002.
53. Kohavi R, Sommerfield D, Feature subset selection using the wrapper method: Overfitting and dynamic search space topology, in *KDD*, pp. 192–197, 1995.

Rishika Sen received her Bachelor of Science (major in Computer Science) and Master of Science degrees in Computer Science from University of Calcutta, India, in the years 2012 and 2014, respectively. She is currently working toward the Doctorate Degree at the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her current research interest includes bioinformatics, computational biology and machine learning.

Losiana Nayak received her Ph.D. degree in Biophysics, Molecular Biology and Bioinformatics from the University of Calcutta, Kolkata, India in the year 2013. She is currently doing research as a UGC Post-Doctoral Fellow (Women) at the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her current research interests include Bioinformatics, Neuroinformatics and Pathogen Informatics.

Rajat K. De is a Professor of the Indian Statistical Institute, Kolkata, India. He obtained his Ph.D. degree from the same institute in the year 2000. He was a Distinguished Postdoctoral Fellow at the Whitaker Biomedical Engineering Institute, the Johns Hopkins University, USA, during 2002–2003. He visited the Department of Medicine, the University of California, San Diego, USA in the years 2017 and 2018, with a Fulbright-Nehru Academic and Professional Excellence Fellowship. During the last 15 years, Professor De has been working in the area of bioinformatics and in silico systems biology. Recently, he has started working on Big Data Analytics and Deep Learning in the domain of bioinformatics and systems biology. He has about 90 research papers published in international journals, conference proceedings and in edited books, and co-edited three books.