**Supplementary Methods**

**1.1 Sequence features**

All the above human protein sequences were downloaded from the UniProt database. For the prediction of proteins and PPIs, the sequence features, such as amino acid composition (AAC), dipeptide composition (DC), pseudo-amino acid composition (PAAC) and composition-transition-distribution (CTD) were earlier reported as important features [[1-3](#_ENREF_1)]. We computed AAC, DC, PAAC and CTD using PyDPI, a freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies [[4](#_ENREF_4)]. We used these sequence features to discriminate the bacterial from viral targeted human proteins.

**1.2 Network features**

To compute network features for human proteins, we retrieved expert-curated human PPIs from the Human Protein Reference Database (HPRD) (Release 9) [[5](#_ENREF_5)] and constructed a network using these PPIs. Network analyzer (cytoscape plugin) is used to compute the network properties, such as degree, closeness centrality, neighborhood connectivity, average shortest path length, betweenness centrality, clustering coefficient, topological coefficient, eccentricity, and radiality [[6](#_ENREF_6)].

**1.3 Gene Ontology (GO) features**

All the GO identifiers (IDs) for the 1,618 and 3,916 bacterial and viral targeted human proteins, respectively were downloaded from UniProt. We found a total of 23,737 GO IDs for 1,618 bacteria targeted human proteins, while the same for the viral targeted human proteins were 67,035. The occurrence of each GO ID was counted separately for the above two groups, followed by sorting based on the occurrence value. The top 100 and 280 GO IDs for the bacterial and viral targeted human proteins were extracted for GO features. However, only 282 were unique among the top 380 GO IDs (Supplementary Table 3). Therefore, we considered the unique IDs for GO features (Supplementary Figure SF1). For each human protein, the presence or absence of the top GO ID was considered as 1 or 0, respectively.

**1.4 Classification**

The distinction between the bacterial and viral targeted human proteins may be viewed as a binary (two-class) classification problem. To differentiate between the proteins, we used well-known classifiers, such as SVM, RF, and DNN.

**1.4.1 Support Vector Machines (SVM):** SVM classifier explicitly maps the data over a vector space to find a decision surface that maximizes the margin between data points of two classes. For the SVM classifier, we used the scikit-learn python package [[7](#_ENREF_7)]. To find the best performance of the SVM classifier, we tested different combinations of cost and gamma parameters of radial basis function (RBF).

**1.4.2 Random Forest (RF):** Several Decision Trees (DTs) grow simultaneously using a random subset of features in RF. In the RF classifier, each tree is a new object and "votes" for that class. Based on a majority vote, the forest elects the classification. We also used the scikit-learn python package for the RF classifier. Optimal parameters were utilized to find the best performance.

**1.4.3 Deep Neural Networks (DNN):** DNN method was shown to perform well with diverse problems. DNN is more robust and useful than other methods for complex classification problems and is becoming a popular algorithm in the field of modern computational biology. We used TensorFlow DNN, which is a widely-used deep learning package for classification to discriminate the bacterial and viral targeted human proteins [[8](#_ENREF_8)].

**1.5 10-fold cross-validation**

To avoid the performance bias of the prediction methods, we used the 10-fold cross-validation technique. In 10-fold cross-validation, the whole dataset is divided into 10 sets (folds) of equal or nearly equal sizes. Training and testing are repeated 10 times so that each time, a different set (fold) goes out for testing, while the remaining 9 sets (folds) are used for training. The average performance measures over the 10 folds are considered for the overall performance of the model.

**1.6 Feature selection**

We used several feature selection methods, such as univariate feature selection (UFS), recursive feature elimination (RFE), feature selection using SelectFromModel (SFM) and tree-based feature selection (TBFS). In UFS, the K best features were selected based on the univariate statistical tests. We used all the univariate statistical test methods available in scikit-learn for the purpose of classification. In RFE, the least important features are excluded in each recursive step, until the desired number of features is reached. The important features are selected from the model in SFM. In TBFS, tree-based estimator computes the importance of the features and irrelevant features are discarded.

**1.7 Performance measures**

The performance measures of the classification problem, such as sensitivity, specificity, accuracy, positive predictive value (PPV or precision), Mathew’s correlation coefficient (MCC) and F1 score were calculated using the following equations:









***Where,***

***True Positive (TP): Bacterial targeted human proteins are correctly identified as bacterial targeted human proteins.***

***False Positive (FP): Viral targeted human proteins are incorrectly identified as bacterial targeted human proteins.***

***True Negative (TN): Viral targeted human proteins are correctly identified as viral targeted human proteins.***

***False Negative (FN): Bacterial targeted human proteins are incorrectly identified as viral targeted human proteins.***

The area under the receiver operating characteristic curve (AUC) for all the cases, were also computed.

**1.8 GO enrichment analysis**

The top 100 bacterial targeted and the same number of viral targeted human proteins predicted by our method, were considered for GO enrichment analysis. To this end, we used Enrichr, a comprehensive gene set enrichment analysis web server, 2016 update [[9](#_ENREF_9)]. We considered only the biological process terms with P-values < 0.05 for the GO enrichment analysis.

**1.9 Pathway enrichment analysis**

The above mentioned 200 human proteins (100 each of the bacterial and viral targeted proteins) were also considered for pathway enrichment analysis. We used the Reactome Pathway Knowledgebase for this purpose [[10](#_ENREF_10)]. Pathways with P-value < 0.05 were treated as enriched pathways.

**References**

[1] P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. I. Siu, “AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest,” *Sci Rep,* vol. 8, no. 1, pp. 1697, Jan 26, 2018.

[2] P. K. Meher, T. K. Sahu, A. Banchariya, and A. R. Rao, “DIRProt: a computational approach for discriminating insecticide resistant proteins from non-resistant proteins,” *BMC Bioinformatics,* vol. 18, no. 1, pp. 190, Mar 24, 2017.

[3] P. K. Meher, T. K. Sahu, J. Mohanty, S. Gahoi, S. Purru, M. Grover, and A. R. Rao, “nifPred: Proteome-Wide Identification and Categorization of Nitrogen-Fixation Proteins of Diaztrophs Based on Composition-Transition-Distribution Features Using Support Vector Machine,” *Front Microbiol,* vol. 9, pp. 1100, 2018.

[4] D. S. Cao, Y. Z. Liang, J. Yan, G. S. Tan, Q. S. Xu, and S. Liu, “PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies,” *J Chem Inf Model,* vol. 53, no. 11, pp. 3086-96, Nov 25, 2013.

[5] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, “Human Protein Reference Database--2009 update,” *Nucleic Acids Res,* vol. 37, no. Database issue, pp. D767-72, Jan, 2009.

[6] Y. Assenov, F. Ramirez, S. E. Schelhorn, T. Lengauer, and M. Albrecht, “Computing topological parameters of biological networks,” *Bioinformatics,* vol. 24, no. 2, pp. 282-4, Jan 15, 2008.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[8] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: A system for large-scale machine learning,” *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265--283, 2016.

[9] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic Acids Res,* vol. 44, no. W1, pp. W90-7, Jul 8, 2016.

[10] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, “The Reactome Pathway Knowledgebase,” *Nucleic Acids Res,* vol. 46, no. D1, pp. D649-D655, Jan 4, 2018.