Accepted Manuscript

DRUG DISCOVERY
TODAY
ELSEVIER
Authoritative reviews and expert opinion
Industry focused, peer reviewed
Special issue on age-related sensory impairment
Produced in association with RNID and Fight-for-Sight
• Screening for drugs that prevent and cause hearing loss
• Delivery of anti-angiogenic therapies for retinal disease
• Treatment strategies for age-related optic nerve disease
www.drugdiscoverytoday.com
RNID
FIGHT FOR SIGHT

Title: DrugMiner: comparative analysis of machine-learning algorithms for prediction of potential druggable proteins

Author: Ali Akbar Jamali Reza Ferdousi Saeid Razzaghi Jiuyong Li Reza Safdari Esmaeil Ebrahimie

PII:	S1359-6446(16)00027-1
DOI:	http://dx.doi.org/doi:10.1016/j.drudis.2016.01.007
Reference:	DRUDIS 1738

To appear in:

Received date:	15-11-2015
Revised date:	5-12-2015
Accepted date:	19-1-2016

Highlights:

-Developing a Novel Machine Learning Tool for Prediction of Potential Druggable Proteins

-Remarkable High Performance of Employed Models in Prediction

DrugMiner[LM1]: comparative analysis of machine-learning algorithms for prediction of potential druggable proteins

Ali Akbar Jamali[1], Reza Ferdousi[2], Saeid Razzaghi[3], Jiuyong Li[4], Reza Safdari[2], and Esmaeil Ebrahimie[4,5,6]

[1]Department of Bioinformatics, Research Institute of Modern Biological Techniques (RIMBT), The University of Zanjan, Zanjan, Iran
[2]Department of Health Information Management and School of Allied-Health Sciences, Tehran University of Medical Sciences, Tehran, Iran
[3]Department of Computer Engineering, Faculty of Engineering, The University of Zanjan, Zanjan, Iran
[4]School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, The University of South Australia, Adelaide, SA, Australia
[5]Department of Genetics & Evolution, School of Biological Sciences, The University of Adelaide, Adelaide, SA, Australia
[6]School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, SA, Australia
Corresponding authors: Safdari, R. (rsafdari@tums.ac.ir); Ebrahimie, E. (Esmaeil.Ebrahimie@unisa.edu.au)

Application of computational methods in drug discovery has received increased attention in recent years as a way to accelerate drug target prediction. Based on 443 sequence-derived protein features, we applied the most commonly used machine-learning methods to predict whether a protein is druggable as well as to opt for superior algorithm in this task[LM2]. In addition, feature selection procedures were used to provide the best performance of each classifier. When run on all features, Neural Network was the best classifier, with 89.98% accuracy, based on a k-fold cross-validation test. Among all the algorithms applied, the optimum number of most-relevant features was 130, according to the Support Vector Machine-Random Forest (SVM-RF) algorithm. This study resulted in the discovery of new drug target in, among others, cell signaling pathways, gene expression, and signal transduction. The DrugMiner web tool was developed based on the findings of this study to provide researchers with the ability to predict druggable proteins. DrugMiner is freely available at www.DrugMiner.org.

Keywords: machine learning; accuracy; drug target; druggable; feature.

Introduction

Following completion of the Human Genome Project, there is the opportunity for pharmacologists to develop new drugs with specific targets in disease. Despite this promise, there have been few new drugs brought to market. Not only is the number of newly designed and approved drugs small, but their effectiveness in treating disease is also poor. The main reason for this is that the systems biology of most diseases is extremely complicated, making the drug–target network intricate and difficult to navigate. Thus, it is an arduous job to design unique but efficacious drugs for many diseases [1].

Analysis of the entire human genome resulted in the identification of many potential drug targets. Target-based drug discovery is a commonly used technique because it can reduce the costs of some laboratory experiments. Appropriate drug targets should have functional characteristics, such as druggability, and must be involved in significant biological pathways [2]. Therefore, it is crucial to continue to identify potential targets for drug design for the discovery of new therapeutic agents.

A peptide, protein, or nucleic acid can be considered druggable if it interacts with a drug or nutraceutical molecule. Based on recent studies, most druggable proteins are classified as functional proteins (enzymes and hormones), ion channels, transporters, G-protein-coupled receptors (GPCRs), and nuclear receptors. Enzymes and GPCRs are the most significant target proteins [3] and more than half of drug targets are categorized under only two protein families: ‘receptors’ and ‘ion channels’. These drug targets mostly comprise GPCRs (26.8%), nuclear receptors (13.0%), and voltage-gated (5.5%) and ligand-gated (7.9%) ion channels [4].

In recent years, progress in drug discovery has benefitted from the application of computerized modeling and algorithms in biology, leading to the emergence of ‘computational drug discovery’ [5,6]. This term encompasses all areas and tools that can be applied to design and discover new drugs in complex diseases systems [7,8]. In such systems, drug–disease networks can be constructed to clarify how drugs affect targets in single and/or multiple diseases [9–11]. The high-throughput nature of computational approaches can complement the output of traditional experimental techniques [12].

Computational approaches in drug target discovery vary from statistical to machine-learning methods. For example, some researchers have applied functional domains and the secondary structure of proteins to analyze potential targets for drugs [13]. Others have analyzed binding sites to predict whether drug or drug-like compounds can bind on the surface of a protein according to its 3D structural properties [14,15]. These approaches depend heavily on the availability of protein 3D structures but, given the lack of these, their application is limited. Recently, machine-learning approaches have been used to predict drug target proteins. In these approaches, simple sequence properties, such as amino acid and di-peptide content and/or frequency, are used to predict potential targets [16–18]. Computationally calculated structural amino acid and/or protein features are useful because they can be easily calculated based on sequence and frequently predict protein function accurately [19–26].

In this study, we applied different machine-learning algorithms to make multi-step predictions on various structural and functional features of proteins. We performed a comparative analysis of machine-learning algorithms to determine which classifier(s) predicted druggable proteins with appropriate performance in terms of their accuracy, sensitivity, and specificity.

Materials and methods

Data acquisition and preparation

Prediction process needs two types of data: training and testing data. The training data set includes both positive (proteins that can interact with drugs) and negative (proteins that cannot be considered as drug target) samples [24]. The 1611 human proteins approved as drug targets and stored in the DrugBank database were retrieved (see Appendix 1 in the supplementary information online). Among these proteins, there were some sequences that were similar in terms of their sequence content and features, which resulted in misclassification and noise in prediction, as well as increased time for algorithm implementation. To overcome these issues, similar sequences were removed. Finally, 1224 protein sequences were selected to comprise the positive sample set (see Appendix 1 in the supplementary information online). Given that there is no available database for proteins that have been recognized as nondrug targets, the methods proposed by Bakheet and Doig, and by Li and Lai [18,27] were used to establish the negative set. As a first step, human-origin protein sequences from Swiss-Prot were retrieved. Then, all 1611 drug targets and all their related families based on sequence similarity and homology from the Pfam database (Pfam 27.0, March 2013, 14 831 families) were eliminated. In addition, research and experimental proteins known as drug targets of human origin from DrugBank [28,29], Therapeutic Target Database (TTD) [30], and members of their related families were also omitted. To achieve a better outcome, identical records of proteins were also eliminated. This approach resulted in 1319 protein sequences considered as nondrug targets representing the negative data set (see Appendix 1 in the supplementary information online). Importantly, the proteins that remained [LM3]were not validated as nondrug target proteins and novel targets might still be found in this data set. Given that only 10–15% of human proteins are druggable [31], this defeat is small [LM4]and would be unlikely to affect the prediction results.

The sequences of all the proteins selected, representing 142 483 known human sequences, were retrieved from the Uniprot database [32] and used in the prediction algorithm.

Attribute extraction and selection

It is well established that different attributes of proteins have a significant role in the interaction between proteins or between one or more drugs and a protein. These attributes vary from sequence-based features to physicochemical features [19,21,22,24–26,33,34]. In this study, attributes were grouped into three categories: (i) Group 1: physicochemical properties of protein sequences, based on amino acid values in terms of their physicochemical parameters, such as length, weight, hydrophobicity, alpha helix, and so on; (ii) Group 2: amino acid composition, which was calculated based on the frequency of the 20 amino acid residues in the protein sequences; and (iii) Group 3: dipeptide composition, which was established based on the frequency of amino acid dimers in the protein sequences (Table 1).

There were 23, 20, and 400 attributes in groups 1, 2, and 3, respectively, all of which have the potential to affect the performance of an algorithm. However, the occurrence of abundant and redundant attributes leads to dimensionality, increasing the time for machine training and limiting the generalization of the developed model. For the prediction of target proteins, we tested our algorithms on single groups of attributes, different combinations of groups, as well as on a combination of all three groups of attributes (Figure 1). Furthermore, to identify effective features [LM5]in the drug–protein interaction, we applied two feature selection algorithms: SVM and Relief were used to identify and mine the 443 attributes that were significant in distinguishing negative and positive classes. Classification algorithms were separately tested with an increasing number of weighed attributes by both feature selection algorithms to find one with the superior performance (see Appendix 2 in the supplementary information online).

Implemented machine-learning predictors

Here, we evaluated different machine-learning approaches to determine which one predicted drug targets with the highest performance. To predict whether a protein could be considered druggable, six of the most well-known prediction algorithms [SVM, Neural Network (NN), k-Nearest Neighborhood (kNN), Naïve Bayes, RF, and Decision Tree (DT)] were used to develop our predictor. The hypotheses underlying each of these algorithms were the minimization of empirical risk and reducing the errors in the training set. Parameters of these algorithms are summarized in Appendix 3 in the supplementary information online. These classifiers were chosen and tested with Orange data mining software (http://orange.biolab.si) and RapidMiner [35], as previously described [34,36].

Prediction assessment

A standard assessment method was essential to evaluate the performance of each algorithm. To do this, we used a k-fold cross-validation in which 80% of the data set was used as a training set and the remaining 20% was used as a test set. The performance of the algorithms was evaluated in terms of gold standards, as detailed below. These parameters were computed based on the values of true negatives (TN), true positives (TP), false positives (FP), and false negatives (FN).

- Accuracy (ACC): percentage of target and nontarget proteins correctly predicted:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \times 100$$

- Sensitivity (SN): percentage of drug target proteins that were predicted correctly:

$$SN = \frac{TP}{TP+FN} \times 100$$

- Specificity (SP): percentage of nontarget proteins that were correctly predicted:

$$SP = \frac{TN}{TN+FP} \times 100$$

- Area under the curve (AUC): this parameter is a logical evaluation for model performance. Its value ranges from 0 to 1, where 1 represents the best performance and 0 is the worst performance. AUC = 0.5 when random ranking is used.
- Matthews Correlation Coefficient (MCC): this value ranges from –1 for worst prediction to +1 for accurate prediction; zero indicates random prediction:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \times 100$$

- Precision or positive predictive value (PPV) = $PPV = \frac{TP}{TP+FN} \times 100$
- Negative predictive value (NPV) = $NPV = \frac{TN}{TN+FN} \times 100$
- F-measure: this parameter is a combined evaluation of precision and recall:

$$F - \text{measure} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$$

Predictor implement

Finally, the algorithm that had the best performance was selected as the predictor to be implemented over all available human proteins (142 483 entries) to discover which proteins could be considered as drug targets. The output for this step resulted in 10 091 potential targets, which are freely available at www.DrugMiner.org.

Figure 1 summarizes the approach used her for the prediction of druggable proteins.

Target class

The gene ontology of the 10 091 proteins identified as possible drug targets was examined using the Comparative GO web tool [37,38].

Results

Attribute selection

As mentioned above, the quantity and quality of attribute sets affected the classifier performance; thus, it is essential to select the optimum number of these attributes. Classifiers were tested with different sets of feature groups (see Appendix 4 in the supplementary information online for complete results). Of the seven possible group subsets, the subset in which all the three feature groups occurred demonstrated the best performance. Table 3 summarizes the results for classifiers when all features were applied.

Additionally, the SVM and Relief feature selection algorithms were used to select the optimal number of attributes. According to the performance of the classifiers with different numbers of features, the optimum number of features that were introduced by the two feature selection algorithms were 130 and 180 for SVM and Relief, respectively (see Appendix 2 in the supplementary information online). In addition, between these two peaks, 130 features of SVM showed the best performance. The performance of classification was evaluated after attribute selection. The 92.10% ACC, 92.80% sensitivity, 91.34% specificity, 84.17% (MCC), and 92.03% precision seem to be logical than those of 443 attributes [LM6]where these values were 89.98%, 90.14%, 89.38%, 79.82%, and 90.14%, respectively. Table 2 summarizes the values for the performance of the classifiers when features were weighed by SVM.

Classifier selection

Six machine-learning approaches were used to develop a classifier to distinguish between target and nontarget proteins in the human proteome. The performance of each algorithm, using all attributes, is summarized in Table 3. Generally, the overall ACC was dominant parameter used to observe the predictive power of a model [18]. It was clear that the best classification performance belonged to the NN classifier and it seemed logical to set our predictor based on this classifier. The feature set comprising all feature groups resulted in the best discrimination between target and nontarget proteins, with an accuracy of 89.78%, sensitivity of 90.14%, and specificity of 89.38% (Table 3, Figure 2).

Target class

Gene Ontology analyses of the 10 091 predicted targets indicated that the targets belonged to protein subclasses that have significant roles in biological processes including cell signaling pathways, gene expression, and signal transduction. These proteins functionally act as enzymes, cell surface receptors, inhibitors, activators, and carriers, and most of the drug targets were classified as enzymes or hormones. Thus, the targets predicted by DrugMiner provide novel candidates that now require laboratory validation.

Discussion

In this study, we used the six most common machine-learning algorithms to develop a prediction model for proteins that can be used as drug targets. The performance of each algorithm was determined by evaluating how correctly they could predict whether proteins were target or nontarget. The performance metrics included the rate of accuracy, sensitivity, and specificity of each algorithm.

Sequence properties determine whether a protein is targetable. Generally, the application of more relevant attributes is key to designing a simple and functional model that has better prediction performance [33,34]. Previous studies related to druggable protein predictions concentrated on only a few attributes, which are unlikely to represent all aspects of protein–drug interactions and also have issues related to limitations of estimations and assumptions [17,39]. However, in this study, we increased the number of features to find the key ones, which is one of the reasons for the high accuracy of prediction in this study [19,23,24,26].

Among the applied algorithms, NN showed a superior performance to that of the Naïve Bayes, SVM, kNN, RF, and DT models. More recent studies have compared approaches based on machine-learning techniques, such as RF, rotation forest, and SVM. However, in the current study, the performance of different classifiers was compared, and selected for on the basis of the best predictor of drug targets.

It is important to determine which aspects of the amino acids of a protein determines their potential as a drug target.. It has been proposed that the molecular function of amino acid residues has an important role in protein–drug interaction [40]. For instance, the structures of binding sites have been analyzed more generally in terms of their general residue preferences. There is a close link between these functional groups and DrugMiner because the latter was developed based on the features reported to be effective properties of residues, such as the charge of the protein [41–43]. Our selection proved the importance of this feature because it was ranked as a high-weighted feature. Many most features, including amino acid hydrophobicity, acidity, alpha helix, and the existence of sulfur atoms in protein structures to establish disulfide bonds, also participate in protein–drug interactions [44–46]. These features and their role in the druggability of proteins, guarantee the reliability of DrugMiner in the identification of druggable proteins because it weighed these features as pivotal features (see Appendix 2 in the supplementary information online).

Altogether, this study demonstrates that combination of different subsets of protein attributes and efficient machine-learning algorithms can remarkably improve and boost the predictability of target proteins. Thus,, results of this study could be used by researchers in studies of drug–protein interactions .

Concluding remarks

The application of computational approaches in the prediction of human drug targets can boost drug design and drug discovery. This study compared machine-learning methods running on sequence-based protein features among three groups of proteins to find possible targets or nontarget for drugs. Target proteins have particular features that govern their suitability for interactions with drugs [3,47]. Based on these characteristics, the NN classifier was established and used to predict potential targets for drugs among all the proteins in our study. The results showed that NN is efficient in predicting those proteins that are more likely to be functional as drug targets. A web tool, DrugMiner (www.DrugMiner.org), was developed based on the results of this study to aid researchers in predicting druggable proteins.

References

1 Lindsay, M.A. (2005) Finding new drug targets in the 21st century. *Drug Discov. Today* 10, 1683–1687

2 Sams-Dodd, F. (2005) Target-based drug discovery: is something wrong? *Drug Discov. Today* 10, 139–147

3 Zheng, C.J. *et al.* (2006) Therapeutic Targets: Progress of Their Exploration and Investigation of Their Characteristics. *Pharm. Rev.* 58, 259–279

4 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat Rev. Drug Discov.* 5, 993–996

5 Ou-Yang, S.-s. *et al.* (2012) Computational drug discovery. *Acta Pharm. Sin.* 33, 1131–1140

6 Sliwoski, G. *et al.* (2014) Computational Methods in Drug Discovery. *Pharm. Rev.* 66, 334–395

7 Materi, W. and Wishart, D.S. (2007) Computational systems biology in drug discovery and development: methods and applications. *Drug Discov. Today* 12, 295–303

8 Matter, H. (2004) Computational Medicinal Chemistry for Drug Discovery. *Drug Discov. Today* 9, 350

9 Berg, E.L. (2014) Systems biology in drug discovery and development. *Drug Discov. Today* 19, 113–125

10 Margineanu, D.G. (2014) Systems biology, complexity, and the impact on antiepileptic drug discovery. *Epilepsy Behav.* 38 , 131–142

11 Yamanishi, Y. *et al.* (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254

12 Lipinski, C.A. *et al.* (2012) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 64, 4–17

13 Ahmadi Adl, A. *et al.* (2012) Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences. *J. Biomol. Struct. Dynamics* 29, 1127–1137

14 Kinnings, S.L. *et al.* (2009) Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* 5, e1000423

15 Xie, L. *et al.* (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.* 5, e1000387

16 Huang, C. *et al.* (2010) Predict potential drug targets from the ion channel proteins based on SVM. *J. Theor. Biol.* 262, 750–756

17 Han, L.Y. *et al.* (2007) Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov. Today* 12, 304–313

18 Li, Q. and Lai, L. (2007) Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics* 8, 353

19 Ashrafi, E. *et al.* (2011) Amino acid features of P1B-ATPase heavy metal transporters enabling small numbers of organisms to cope with heavy metal pollution. *Bioinform. Biol. Insights* 5, 59

20 Ebrahimi, M. *et al.* (2010) Are there any differences between features of proteins expressed in malignant and benign breast cancers? *J. Res. Med. Sci.* 15, 299

21 Ebrahimi, M. *et al.* (2011) Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PLoS ONE* 6, e23146

22 Tahrokh, E. *et al.* (2011) Comparative study of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms. *Genes Genomics* 33, 565–575

23 Zinati, Z. *et al.* (2014) New layers in understanding and predicting α-linolenic acid content in plants using amino acid characteristics of omega-3 fatty acid desaturase. *Comput. Biol. Med.* 54, 14–23

24 Bakhtiarizadeh, M.R. *et al.* (2014) Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. *J. Theor. Biol.* 356, 213–222

25 DELAVARI, A. *et al.* (2014) Determining the structural amino acid attributes which are important in both protein thermostability and alkalophilicity: a case study on xylanase. *BioTechnologia* 95, 161–173

26 KayvanJoo, A.H. *et al.* (2014) Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. *BMC Res. Notes* 7, 565

27 Bakheet, T.M. and Doig, A.J. (2009) Properties and identification of human protein drug targets. *Bioinformatics* 25, 451–457

28 Law, V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42 (D1), D1091–D1097

29 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36 (Suppl. 1), D901–D906

30 Zhu, F. *et al.* (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 38 (Suppl. 1), D787–D791

31 Thangudu, R.R. *et al.* (2012) Modulating protein–protein interactions with small molecules: the importance of binding hotspots. *J. Mol. Biol.* 415, 443–453

32 Consortium, T.U. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212

33 Nguyen, M. *et al.* (2007) Amino acid features for prediction of protein-protein interface residues with support vector machines. *Evol. Comput. Machine Learn. Data Mining Bioinformat.* 4447, 187–196

34 Ebrahimie, E. *et al.* (2011) Protein attributes contribute to halo-stability, bioinformatics approach. *Saline Syst.* 7, 1

35 Hofmann, M. and Klinkenberg, R. (2013) *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, CRC Press

36 Ebrahimi, M. *et al.* (2014) Understanding the underlying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein. *PLoS ONE* 9, e96984

37 Fruzangohar, M. *et al.* (2014) Application of global transcriptome data in gene ontology classification and construction of a gene ontology interaction network. *bioRxiv* 004911

38 Fruzangohar, M. *et al.* (2013) Comparative GO: a web application for comparative gene ontology and gene ontology–based gene selection in bacteria. *PLoS ONE* 8, e58759

39 Hajduk, P.J. *et al.* (2005) Predicting protein druggability. *Drug Discov. Today* 10, 1675–1682

40 López-Romero, P. *et al.* (2004) Prediction of functional sites in proteins by evolutionary methods. In *Methods in Proteome and Protein Analysis* (Kamp, R. *et al.*, eds.), pp. 319–340, Springer

41 Law, M.J. *et al.* (2006) The role of positively charged amino acids and electrostatic interactions in the complex of U1A protein and U1 hairpin II RNA. *Nucleic Acids Res.* 34, 275–285

42 Reif, M.M. *et al.* (2012) New interaction parameters for charged amino acid side chains in the GROMOS force field. *J. Chem. Theory Comput.* 8, 3705–3723

43 Kahlen, J. *et al.* (2014) Interaction of charged amino-acid side chains with ions: an optimization strategy for classical force fields. *J. Phys. Chem. B* 118, 3960–3972

44 Nath Jha, A. *et al.* (2010) Amino acid interaction preferences in proteins. *Protein Sci.* 19, 603–616

45 Kresge, N. *et al.* (2008) Amino acid solubility and hydrophobic interactions in proteins: the work of Charles Tanford. *J. Biol. Chem.* 283, e3

46 Vaitheeswaran, S. and Thirumalai, D. (2008) Interactions between amino acid side chains in cylindrical hydrophobic nanopores with applications to peptide stability. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17636–17641

47 Imming, P. *et al.* (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* 5, 821–834

Table 1. The three groups of protein sequence-based features used in this study for drug target prediction

Feature set	Description
Group 1[a]	Length, weight, half-life, alpha helix, beta sheets, isoelectric point Aliphatic index, hydrogen, carbon, nitrogen, oxygen, sulfur Tiny amino acids: (A, C, G, S, T) Small amino acids: (A, C, G, S, T, V, P, D, N) Aliphatic amino acids: (I, L, V) Nonpolar amino acids: (A, C, G, I, L, M, P, V, Y, F, W) Aromatic amino acids: (F, H, Y, W) Polar amino acids: (D, H, E, Q, K, N, T, R, S) Charged amino acids: (H, D, E, R, K) Basic amino acids: (R, K, H) Acidic amino acids: (D, E) Hydrophobic amino acids: (A, C, F, I, L, M, V, Y, W) Hydrophilic amino acids: (D, E, N, Q, K)
Group 2[b]	Amino acid composition of protein sequence
Group 3[c]	Dipeptide composition of protein sequence

[a] 23 features, including physicochemical properties of proteins.

[b]20 features, including frequency of each amino acid in protein structure.

[c]400 features, including frequency of dipeptides in protein structure.

Table 2. Evaluation metrics of NN using features from the SVM feature selection technique[a]

Number of features	Feature									
	SN	**SP**	**MCC**	**ACC**	**AUC**	**PPV**	**FPR**	**NPV**	**FDR**	**F-score**
443	90.14	89.38	79.52	89.78	0.9592	90.14	10.62	89.38	9.86	90.14
430	89.92	89.87	79.77	89.89	0.9608	90.53	10.13	89.21	9.47	90.22
420	90.22	89.38	79.60	89.82	0.9606	90.15	10.62	89.45	9.85	90.19
410	90.37	89.71	80.08	90.05	0.9618	90.44	10.29	89.63	9.56	90.41
400	89.92	90.11	80.01	90.01	0.9624	90.74	9.89	89.24	9.26	90.33
390	90.75	90.11	80.86	90.44	0.9629	90.82	9.89	90.04	9.18	90.78
370	91.05	90.03	81.10	90.56	0.9643	90.78	9.97	90.33	9.22	90.92
350	90.37	90.44	80.79	90.41	0.9652	91.06	9.56	89.71	8.94	90.72
330	91.21	89.71	80.94	90.48	0.9669	90.52	10.29	90.44	9.48	90.86
310	91.13	91.01	82.13	91.07	0.9681	91.62	8.99	90.50	8.38	91.37
290	90.98	90.93	81.89	90.96	0.9698	91.53	9.07	90.34	8.47	91.25
270	91.05	90.77	81.81	90.92	0.9697	91.40	9.23	90.40	8.60	91.23
250	91.96	90.44	82.43	91.23	0.9702	91.20	9.56	91.26	8.80	91.58
230	92.27	91.01	83.30	91.66	0.9715	91.71	8.99	91.61	8.29	91.99
210	91.81	90.93	82.75	91.39	0.9709	91.60	9.07	91.15	8.40	91.71
190	92.12	90.85	82.99	91.51	0.9720	91.56	9.15	91.45	8.44	91.84
170	91.89	91.18	83.07	91.55	0.9723	91.82	8.82	91.25	8.18	91.85
150	92.42	91.01	83.46	91.74	0.9730	91.72	8.99	91.76	8.28	92.07
▸ 130	**92.80**	**91.34**	**84.17**	**92.10**	**0.9728**	**92.03**	**8.66**	**92.17**	**7.97**	**92.41**
110	92.42	91.42	83.85	91.94	0.9727	92.07	8.58	91.80	7.93	92.24
90	91.96	90.52	82.51	91.27	0.9691	91.27	9.48	91.27	8.73	91.62
70	91.05	88.89	79.99	90.01	0.9645	89.83	11.11	90.22	10.17	90.44
50	89.08	87.91	77.00	88.52	0.9579	88.81	12.09	88.20	11.19	88.95
30	85.97	85.21	71.18	85.61	0.9332	86.24	14.79	84.93	13.76	86.10
10	79.30	82.84	62.11	81.01	0.8869	83.28	17.16	78.79	16.72	81.24

[a]The performance of the NN classifier with an increasing number of features ranked by the SVM feature selection. The optimum number of features was 130, where the classifier shows the best performance in terms of: SN (sensitivity), SP (specificity), MCC (Matthew's correlation coefficient), ACC (accuracy), AUC (area under the curve), PPV (precision or positive predictive value), FPR (false positive rate). NPV (positive predictive value), and FDR (false discovery rate).

Table 3. Golden standard values for six machine-learning algorithms where all 443 protein features were applied[a]

Classifier	Feature									
	SN	**SP**	**MCC**	**ACC**	**AUC**	**PPV**	**FPR**	**NPV**	**FDR**	**F-score**
Naïve Bayes	80.44	66.91	47.88	73.93	0.8187	72.37	33.09	76.04	27.63	76.19
KNN	88.86	80.07	69.32	84.62	0.9126	82.77	19.93	86.96	17.23	85.7
▸ NN	**90.14**	**89.38**	**79.52**	**89.78**	**0.9592**	**90.14**	**10.62**	**89.38**	**9.86**	**90.14**
RF	89.54	84.31	74.04	87.02	0.9444	86.02	15.69	88.21	13.98	87.74
SVM	90.07	87.83	77.94	88.99	0.9562	88.86	12.17	89.14	11.14	89.46
DT	76.72	75.08	51.8	75.93	0.7589	76.84	24.92	74.96	23.16	76.78

[a]Abbreviations, NN, k-nearest neighbor; NN, neural network; RF, random forest; SVM support vector machine DT, decision tree; SN, sensitivity; SP, specificity; MCC, Matthew's correlation coefficient; ACC, accuracy; AUC, area under the Curve; PPV, precision or positive predictive value; FPR, false positive rate. NPV, positive predictive value; FDR, false discovery rate.

Figure 1. Illustration of the proposed approach for potential drug target prediction. Feature group 1 (G1) comprised the physicochemical properties of protein sequences; group 2 (G2) included amino acid composition of the proteins, and group3 (G3) involved the dipeptide composition of the proteins.

Figure 2. Area under the curve (AUC) plots of different algorithms: (A) Optimum number of features produced by support vector machine (SVM). (B) All 443 attributes.

Figure 1

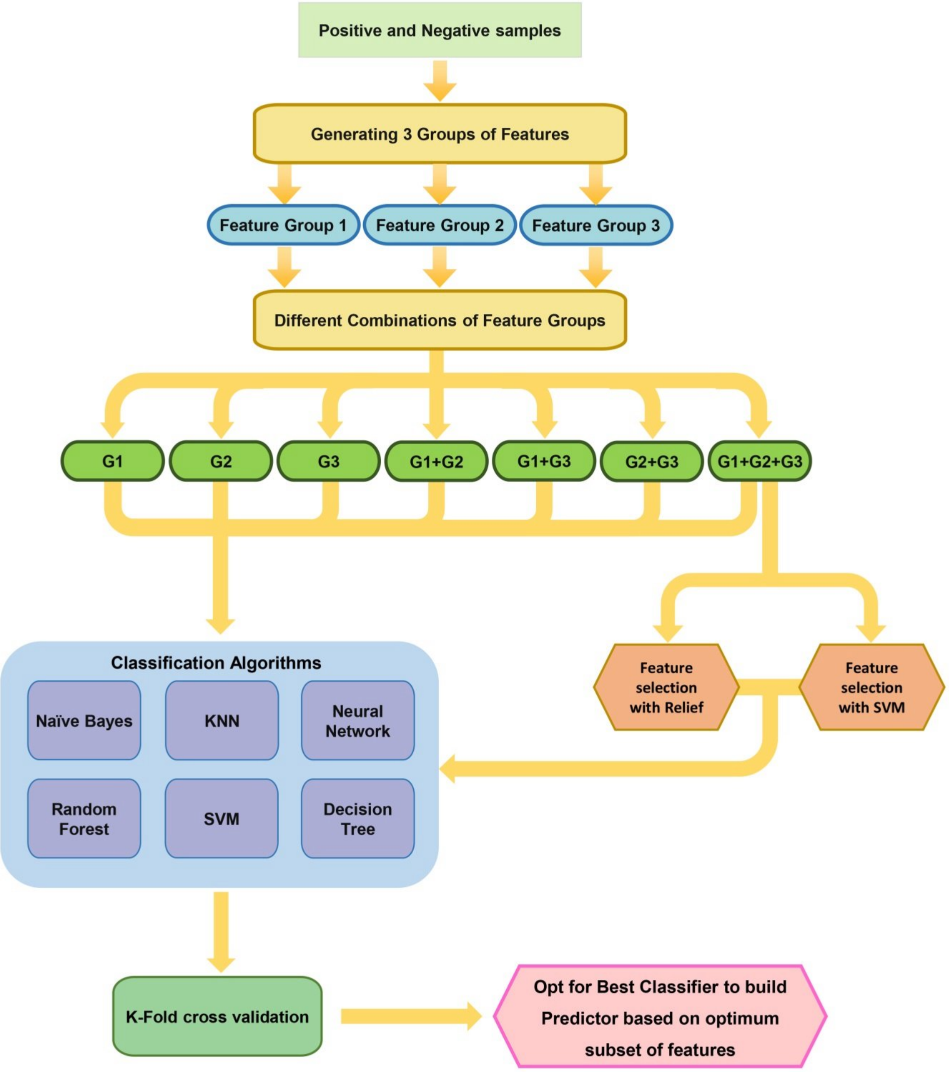
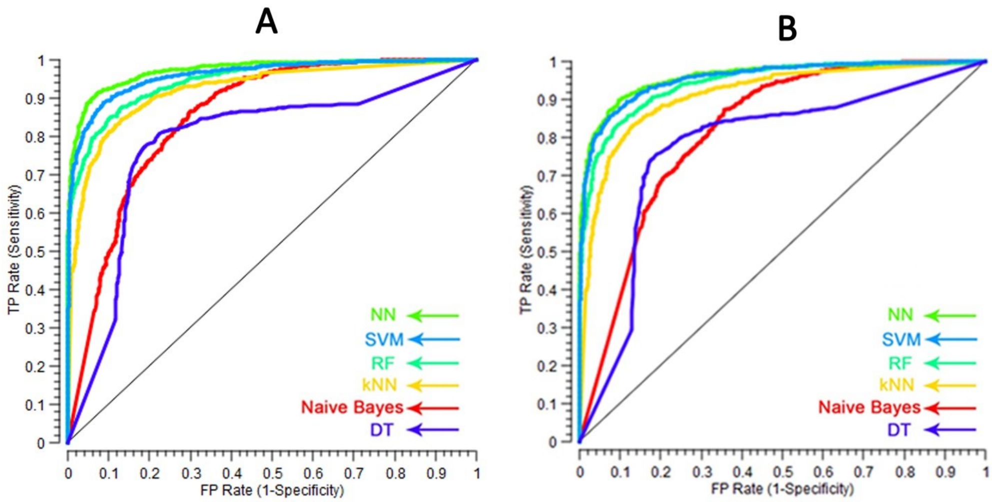

Positive and Negative samples
Generating 3 Groups of Features
Feature Group 1
Feature Group 2
Feature Group 3
Different Combinations of Feature Groups
G1
G2
G3
G1+G2
G1+G3
G2+G3
G1+G2+G3
Classification Algorithms
Naïve Bayes
KNN
Neural Network
Random Forest
SVM
Decision Tree
Feature selection with Relief
Feature selection with SVM
K-Fold cross validation
Opt for Best Classifier to build Predictor based on optimum subset of features

Figure 2


A
B
TP Rate (Sensitivity)
FP Rate (1-Specificity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
NN
SVM
RF
kNN
Naive Bayes
DT