molecules

MDPI

Article

Computational Prediction and Analysis of Associations between Small Molecules and Binding-Associated S-Nitrosylation Sites

Guohua Huang [1,2,*], Jincheng Li [1,2] and Chenglin Zhao [1,2]

[1] Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang 422000, China; hnsyljc63@163.com (J.L.); chenglinzhao@126.com (C.Z.)
[2] College of Information Engineering, Shaoyang University, Shaoyang 422000, China
* Correspondence: guohuahhn@163.com; Tel.: +86-739-540-0925

Received: 24 February 2018; Accepted: 9 April 2018; Published: 19 April 2018

Abstract: Interactions between drugs and proteins occupy a central position during the process of drug discovery and development. Numerous methods have recently been developed for identifying drug–target interactions, but few have been devoted to finding interactions between post-translationally modified proteins and drugs. We presented a machine learning-based method for identifying associations between small molecules and binding-associated S-nitrosylated (SNO-) proteins. Namely, small molecules were encoded by molecular fingerprint, SNO-proteins were encoded by the information entropy-based method, and the random forest was used to train a classifier. Ten-fold and leave-one-out cross validations achieved, respectively, 0.7235 and 0.7490 of the area under a receiver operating characteristic curve. Computational analysis of similarity suggested that SNO-proteins associated with the same drug shared statistically significant similarity, and vice versa. This method and finding are useful to identify drug–SNO associations and further facilitate the discovery and development of SNO-associated drugs.

Keywords: SNO; random forest; fingerprints; information entropy; machine learning

1. Introduction

Disease is one of the most serious threats to the safety of lives. In 2006 alone, for example, 1,685,210 new cancer cases and 595,690 cancer deaths took place in the United States, and cancer was rising to become the second-leading cause of death [1]. Although very complicated, the cause for this phenomenon was grouped into two main factors. For one thing, the disease's pathological process is very complex. Environmental, epigenetic, and lifestyle factors jointly determine the development and progress of disease [2]. The cause of the poorly understood Alzheimer's disease, for example, is believed to be a combination of genetics and history of head injuries, depression, or hypertension [3]. For another, the discovery and development of new drugs could not match the requirement of treating the disease at all, because it is time-consuming and labor-intensive. The cost and the time of developing a new drug was conservatively estimated at an average of more than $ 800 million and at an average of 12 to 15 years, respectively [4]. For drug discovery and development, the first important thing is to find drug targets related to diseases.

Although a large number of computational methods have been proposed to predict drug–target interactions in the past decades [5–19], few were dedicated to such specific types of proteins as post-translationally modified proteins and membrane proteins. Because most of the identified targets were not associated with specific diseases, this limited the discovery and development of drugs to a certain extent.

S-nitrosylation (SNO) is a type of post-translational modification where a nitric oxide is covalently attached to a cysteine residue in proteins. In addition to regulation of protein structure and functions [20], SNO plays an important role both in normal health and in a wide range of human diseases [21,22]. For example, aberrant protein SNO was implicated especially in neurodegenerative diseases, including Alzheimer's and Parkinson's diseases [23–26]. Protein SNO was also reported to be associated with some cancers, including lung cancer [27,28]. Accumulating evidence suggested that SNO-protein is a therapeutic target for some degenerative diseases and cancers [27,29]. This provides a new idea to treat these diseases: to interfere with diseases through the specific drugs to mediate relevant SNO proteins. In the paper, we propose a computational method to identify drug–SNO associations.

2. Experimental Data

Associations between drugs and binding-associated SNO sites were downloaded from the dbPTM database (http://dbPTM.mbc.nctu.edu.tw/), an experimentally validated post-translational modification (PTM) database [30,31]. The PTM sites whose side chains are located within 10 Å of a drug-binding SNO site were viewed as drug binding-associated PTMs [30]. We collected 147 pairs of associations between drugs and binding-associated SNO sites, including 38 such SNO sites located in 36 proteins and 125 drugs. All these 147 pairs of associations were regarded as positive samples. We randomly paired 38 SNO sites and 125 drugs. Any pairs that were not identical to positive samples served as negative ones, where SNO sites were supposed not to be associated with the corresponding drug. We yielded 478 negative samples, of which 147, along with the 147 positive samples, composed the training set (see Supplementary Material S1). 36 SNO-protein sequences were retrieved from the Uniprot database, a comprehensive repository of protein sequences and function annotation [32–36]. SNO-protein sequences were cut into peptides with 21 amino acid residues and centering the SNO-modified cysteine. If peptides with the SNO-modified cysteine were of less than 21 residues, we added one "X" character or more to reach it.

3. Method

As shown in Figure 1, the proposed approach consisted mainly of three steps: encoding, training, and predicting. Peptides and small molecules of drugs were encoded into features by information entropy and molecular fingerprints, respectively. Then, the random forest algorithm [37] was used to train the classifier over the training set. For unknown drug–SNO associations, the trained classifier predicted and outputted the answers "yes" or "no".

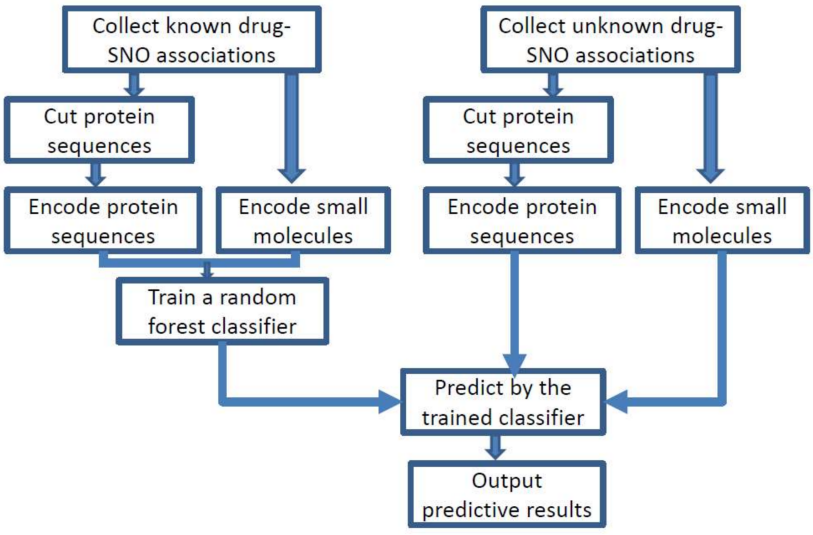

Figure 1. The overview of the proposed method.

3.1. Encoding Small Molecules

Molecular fingerprinting is an important concept in the area of chemoinformatics and bioinformatics, which has been initiated for chemical substructure search [38], and later, widely used for molecular similarity comparison [39], clustering [40], and so forth. Fingerprinting of small molecules is generally a binary string representation. There are generally two main categories: structure-based and hash-based fingerprints. For the former, each bit represents the presence or absence of a specific substructure; namely, 1 standing for presence and 0 for absence, as illustrated in Figure 2A. Different structure-based fingerprint systems depict different aspects of molecular substructure. Electrotopological state (E-state) fingerprints represent the presence/absence of the 79 E-state substructures defined by Kier and Hall in a molecule [41]. The pubchem fingerprint is an 881-bit binary string, covering a wide range of different substructures and features [42,43]. The MACCS fingerprint [44] depicts 166 key substructures including ISOTOPE, ON(C)C, C$=C($A)$A and C=C(C)C (https://list.indiana.edu/sympa/arc/chminf-l/2007-11/msg00058.html), commonly involving most key chemical features for drug discovery and virtual screening [45]. The chemistry development kit (CDK) substructure fingerprint represents information on the presence of 307 substructures. The CDK extended fingerprint [46] extends the CDK fingerprint with additional bits describing ring features. The CDK graph-only fingerprint is a specialized replacement of the CDK fingerprint which does not consider bond orders [46], and Klekota–Roth fingerprints represent the presence/absence of 4860 substructures. The Daylight fingerprint [45] is a type of hash-based fingerprint, which analyzed the molecular fragments of all paths of up to a user-defined number of bonds, and then hashed every one of these paths, as illustrated in Figure 2B. The CDK Daylight fingerprint hashed each atom type, all augmented atoms, and all paths of 2–7 atoms into a 1024-bit binary string. The CDK hybridization fingerprint is a combination of different fingerprints. All fingerprints mentioned above were calculated by the CDK software [47,48] in the webserver ChemDes, a platform providing multiple software in an online manner for computing molecular descriptors and fingerprints [49]. Table 1 listed these fingerprints and the numbers of bits.

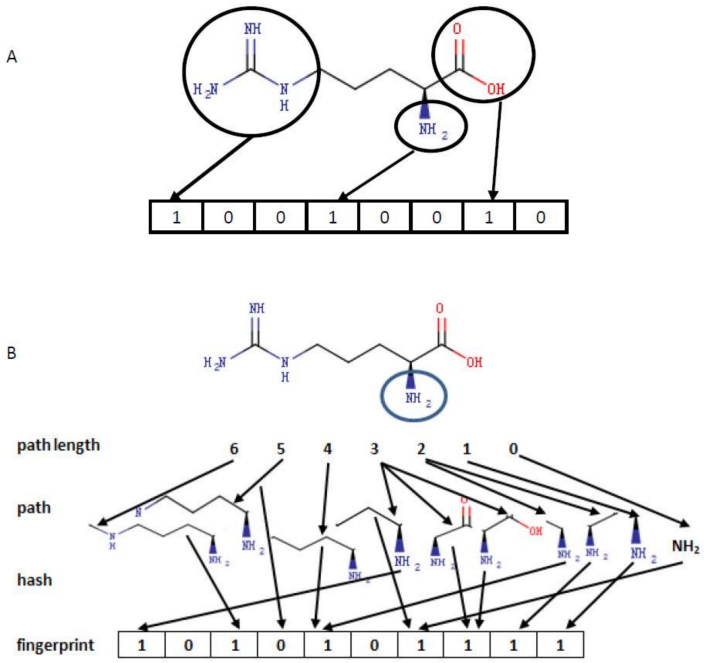

Figure 2. Molecular fingerprints. The top diagram (**A**) illustrates structure-based fingerprints, where eight substructures were explored and three substructures, marked with circles, were present. The bottom diagram (**B**) illustrates hash-based fingerprints, where all paths starting from NH_2 (circle) and of up to a length of six were explored, and each path was then hashed into a binary bit.

Table 1. The categories and the bit numbers of fingerprints.

Category	Number of Bits	Category	Number of Bits
E-state	79	Klekota–Roth	4860
Daylight	1024	MACCs	166
CDK extended	1024	CDK substructure	307
CDK graph	1024	PubChem	881
CDK hybridization	1024		

3.2. Encoding Protein Peptides

In the field of bioinformatics, there are numerous ways of encoding biological sequences [50–56] such as the widely used amino acid composition [52], feature vector [53], pseudo amino acid composition [54], and amino acidphysicochemical properties [55], etc. We used the information entropy to encode each peptide. The information entropy of amino acid (*IEA*) was computed by:

$$IEA(\lambda) = \sum_{i=1}^{n} P_i(\lambda) \log_2 P_i(\lambda) \tag{1}$$

where λ represents one of 21 kinds of amino acid and $P_i(\lambda)$ denotes the probability of occurring at the ith position for the amino acid λ. Equation (1) measures the information on the distribution of position for the specific amino acid. Similarly, the information entropy of the position (*IEP*)was denoted by:

$$IEP(i) = \sum_{\lambda \in \Omega} P_i(\lambda) \log_2 P_i(\lambda) \tag{2}$$

where Ω denotes the set of 21 amino acids. The *IEP* measures information on the probability of an amino acid at a specific position. All these 38 protein peptides in the training set were used to estimate the probabilities both of positions for a specific amino acid and of an amino acid at a specific position. For each peptide, the information entropy over all the peptides minus that over the whole set subtracting this peptide was used to encode it. Because all the residues at the 11th position were always cysteine, its information entropy was 0 and was removed. Finally, we obtained 41 features to depict the peptides, each feature $r'_{i,j}$ normalized by:

$$r_{i,j} = \frac{r'_{i,j} - \max\limits_i \{r'_{i,j}\}}{\max\limits_i \{r'_{i,j}\} - \min\limits_i \{r'_{i,j}\}}, \; j = 1, 2, \cdots, 41 \tag{3}$$

The 41 features represent 21 and 20 information entropies of amino acids and of position, respectively.

3.3. Random Forest

Random forest, which combines bagging and random selection of features, is a type of ensemble learning algorithm [37]. The random forest was described by three key steps: (1) sample with replacement n training examples; (2) randomly select k features; (3) construct a decision tree using n training examples with k features. Repeating the above three steps m times generated m decision trees. For a classification task, the output of testing sample x is majority vote of m decision trees. Due to advantages such as cheap computation time, ability to deal with high-dimensional data, and better performance, the random forest has been widely applied to classification and regression [57–60].

3.3.1. Cross Validation and Metrics

We used ten-fold and leave-one-out cross validations to test the proposed method. In the ten-fold cross validation, the training set was randomly grouped into 10 parts of equal or almost-equal size. Each part was in turn taken as testing samples by the trained classifier over the other nine parts.

This procedure was repeated ten times. The leave-one-out test is an extreme case of the cross-validation test, where each sample is viewed as an independent part. We used the receiver operating characteristic curve (also known as the ROC curve) to depict experimental performances. The ROC curve could be drawn by plotting the true positive rate against the false positive rate at various thresholds. The area under the ROC curve (AUC) was used to quantitatively assess the experimental performance. The AUC ranges from 0 to 1, with 1 and 0.5 representing the best and uninformative performances, respectively.

3.3.2. Computational Environment

All computations were performed on a Microsoft Windows operating system with a 64-bit version (Windows 10) on an Acer personal computer with an Intel® Core™ i5-3210M CPU and 6.0 G RAM.

4. Results and Discussion

The ROC curves of the ten-fold cross validation for nine types of fingerprints were drawn in Figure 3. The MACCS fingerprints performed best, followed by substructures, and then by PubChem fingerprints. Their AUCs were 0.7312, 0.6943, and 0.6508, respectively. The combination of nine fingerprints and the hybridization fingerprint performed worst and second-worst, with their AUCs of 0.3012 and 0.4767, respectively. We repeated the ten-fold cross validation ten times. The mean value is 0.7235. Supposing that AUC was the normal distribution with unknown variance values, the 95% confidence interval of the mean AUC was estimated at [0.7160, 0.7311]. To answer the question of if the random forest is a better algorithm for predicting drug–SNO associations, we further compared it with state-of-the-art learning algorithms: C4.5 [61], the naive Bayes classifier [62], the radial basis function network [63], and Bagging [64]. C4.5 is a type of decision tree algorithm [61]. In 2008, C4.5 ranked first in the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining [65]. The naive Bayes classifier [62] is a Bayes' theorem-based statistical learning model, where variables are supposed to be independent of each other and which makes decisions commonly by computing post-probabilities of samples, given specific classes. The radial basis function network [63] is a specific artificial neural network with radial basis functions as the specific activation function. Bagging [64] is an ensemble learning algorithm, which generally makes decisions by voting over some constituent subclassifiers. Due to excellent performance, these four algorithms have widely been applied to such fields as function approximation and classification. Here, they were used as baseline algorithms for comparison. As shown in Figure 4, the random forest performed best in the leave-one-out cross validation, being 0.1 more than Bagging in terms of AUC.

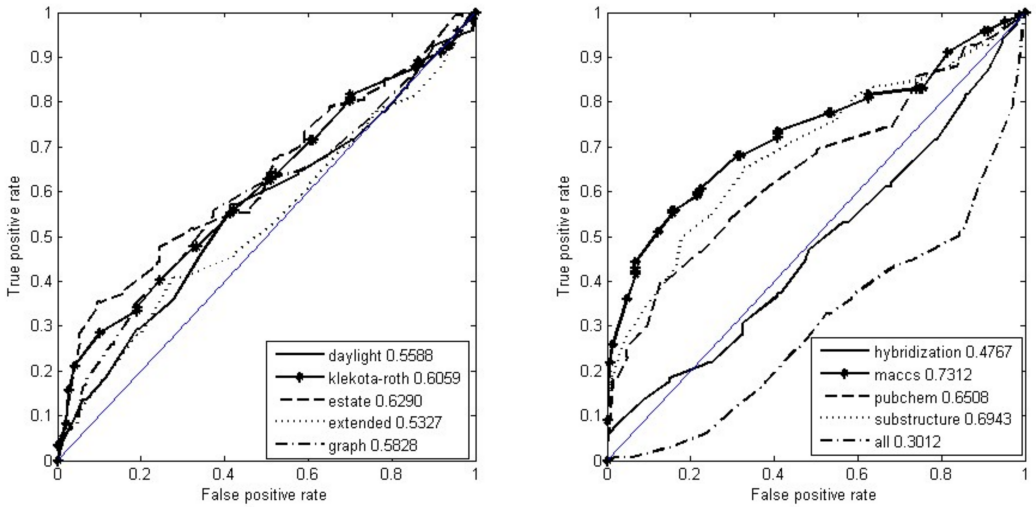

Figure 3. The receiver operating characteristic (ROC) curves of ten types of fingerprint by ten-fold cross validation.

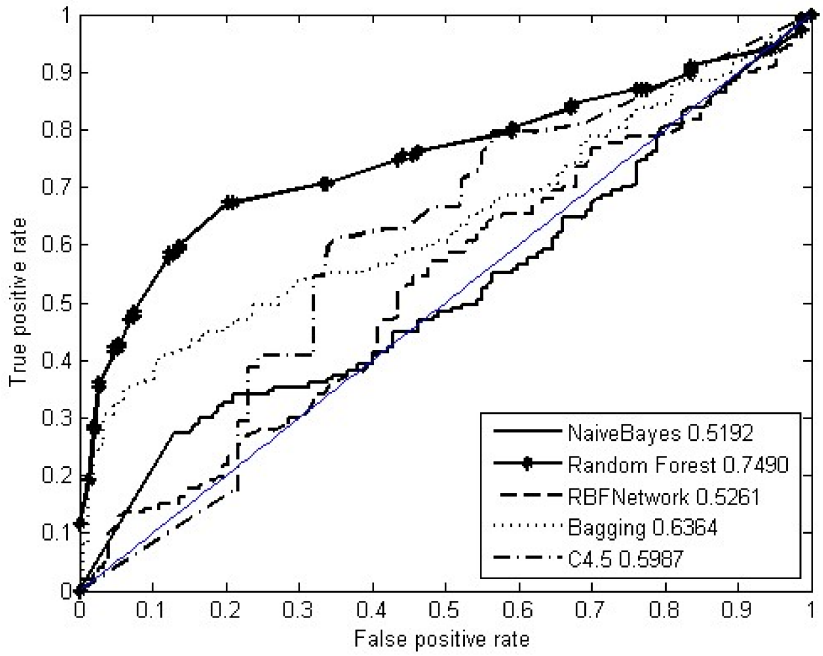

Figure 4. The ROC curves of five algorithms by leave-one-out cross validation.

4.1. Computational Analysis of Associations between Drugs and SNO-Proteins

The common hypothesis in the field of molecular biology is the guide by association (GBA) principle [66]. In order to test whether this hypothesis was applicable to associations between drugs and SNO-proteins, we computed the drug–drug and the protein–protein similarities and compared them. Sequence similarity is a basic concept in the field of molecular biology, on which many hypotheses are based. For example, similar sequences are assumed to possess similar structures, which are in turn assumed to have similar functions. We used sequence similarity to measure protein–protein similarity. Each protein sequence was aligned against the whole 38 protein peptides by the PSI-BLAST program [67]. The matrix $(P_{i,j})_{\ 38 \times 38}$ denoted the alignment score. Each column was normalized by:

$$S(T_i, T_j) = \frac{p_{i,j}}{\max\limits_t p_{t,j}}, i = 1, 2, \cdots, 38, j = 1, 2, \cdots, 38 \tag{4}$$

To keep the similarity matrix symmetrical, let:

$$PS(T_i, T_j) = [S(T_i, T_j) + S(T_j, T_i)]/2 \tag{5}$$

Drug–drug similarity was computed by the Tanimoto coefficient [68], namely:

$$Tc(D_1, D_2) = \frac{|d_1 \cap d_2|}{|d_1| + |d_2| - |d_1 \cap d_2|} \tag{6}$$

where $|A|$ denotes the number of shared fingerprints, and d_1 and d_2 represent fingerprints of the drugs D_1 and D_2, respectively. To keep drug–drug similarity sensible, Tc of less than 0.6 was set to 0.

To test the hypothesis that similar proteins would be associated with similar drugs, we computed similarity PPS^d among proteins $\{p_1, p_2, \cdots p_k\}$ associated with the same drug by:

$$PPS^d = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} TS(p_i, p_j) \tag{7}$$

The similarity PPS_c^d among other proteins not being associated with the drug d was used as the control. We used a permutation test to examine the above hypothesis. The p-value was 0.0020, statistically suggesting that similar SNO-proteins would be associated with the same drugs. Figure 5A demonstrated such a case: two samples P15121-304 and P15121-299 in the protein aldose reductase shared most amino acid peptides, and they were associated with the same drugs: DB02383, DB07030, DB07093, DB07139, DB08084, DB08449, and DB08772 (DBXXXXX is the identifier of a drug in the Drugbank database, which is a unique bioinformatics and cheminformatics resource [69,70]). Similarly, to test hypothesis that similar drugs would be associated with similar SNO-proteins, we computed similarity DDS^p among drugs $\{d_1, d_2, \cdots d_k\}$ being associated with the same protein p by:

$$DDS^p = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Tc(d_i, d_j) \quad (8)$$

The similarity DDS_c^p among other drugs not being linked to the SNO-protein p was used as the control. The p-value in the permutation test is 0.0077, statistically implying that similar drugs would be associated with the same SNO-proteins. As shown in Figure 5B, both drugs phosphoaminophosphonic acid guanylate ester (DB02082) and guanosine-5′-diphosphate (DB04315), which were associated with the same SNO-protein, P63000-178, were very similar in two-dimensional (2D) structure.

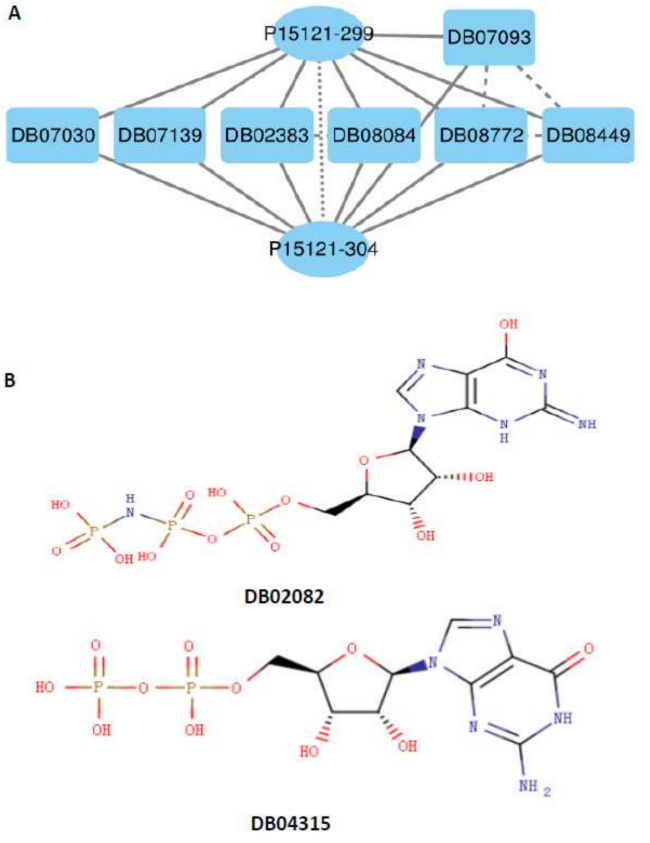

Figure 5. Illustration of cases of similarity. The top diagram (**A**) illustrates the similarity of SNO-proteins associated with the same small molecule. The bottom diagram (**B**) illustrates the similarity of small molecules associated with the same SNO-protein P63000-178.

4.2. Large-Scale Prediction of Unknown Associations of Drugs and SNO-Proteins

In total, 147 of the randomly yielded negative samples in the section "Experimental data" served as negative training samples, and the remaining 331 were used as the testing

samples (see Supplementary Materials S2). Of the 331 random associations, 213 were predicted to be negative by the trained classifier over the training set; its accuracy being viewed as 213/331 = 0.6435. Of 118 predicted positive samples, 12 were with the output of probability of 1, implying that they were potential positive samples. We intended to look for evidence to support the prediction by similarity analysis. Next, we investigated SNO-protein targets of small molecules similar to the studied small molecules (*i.e.*, drugs in the column in Table 2). All the small molecules, except DB07905, have similar drugs which all respectively shared the same SNO-protein as the analogue, as shown in Table 2. For example, the association between the small molecule 3-[[*N*-[4-methyl-piperazinyl]carbonyl]-phenylalaninyl-amino]-5-phenyl-pentane-1-sulfonic acid benzyloxy-amide (DB04427) and the tyrosine-protein phosphatase nonreceptor type 1 (P18031) protein was predicted. The small molecule DB04427 was similar to these small molecules: DB06887, DB07719, DB08003, DB08549, DB08591, DB08593, and DB02827, which all targeted the SNO-protein P18031. Thus, it seemed a rational to identify the association of DB04427 with P18031. The protein aldose reductase (P15121) catalyzes the NADPH-dependent reduction of a wide variety of carbonyl-containing compounds to their corresponding alcohols with a broad range of catalytic efficiencies [32,33]. The small molecule 7*N*-methyl-8-hydroguanosine-5′-diphosphate (DB01960) belongs to a type of organic compounds known as purine ribonucleoside diphosphates. The fact that the small molecule DB01960 was similar to the molecules DB02338 and DB03461, which target the SNO-protein P15121, supported predicted DB01960–P15121 associations to a certain extent. The SNO-protein cathepsin K (P43235) was closely involved in osteoclastic bone resorption and may participate partially in the disorder of bone remodeling. The small molecule L-citrulline (DB00155) was predicted to bind the SNO-protein cathepsin K. L-citrulline shares structural similarity with the small molecules DB04276, DB04523, and DB07592, which were associated with P43235. Thus, it seemed a natural inference to predict the DB00155–P43235 association. We indirectly explained the rationality of the above predicted SNO–drug associations in terms of molecular similarity. Due to limitations of conditions, we did not conduct wet experiments to validate these associations, which will be left for experimental biologists to validate in the coming future.

Table 2. Twelve predicted drug–SNO associations, similar drugs, and SNO-proteins targeted by similar drugs.

Predicted Associations	Similar Small Molecules	SNO-Proteins Targeted by Similar Molecules
DB04427–P18031-215	DB06887, DB07719, DB08003, DB08549, DB08591, DB08593, DB02827	P18031-215
DB01960–P15121-299	DB02338, DB03461	P15121-299
DB04315–P15121-299	DB02338, DB03461, DB08772	P15121-299
DB08213–P18031-215	DB02827, DB06887, DB07134, DB07719, DB07730, DB08003, DB08549, DB08591, DB08593	P18031-215
DB04502–P18031-215	DB01962, DB03483, DB03557, DB06887, DB07719, DB08003, DB08549, DB08591	P18031-215
DB00114–P18031-215	DB01962, DB07480	P18031-215
DB02051–P18031-215	DB06887, DB07719, DB08549, DB08591	P18031-215
DB07905–P18031-215	not existing similar drugs	P18031-215
DB08607–P18031-215	DB02072, DB02827, DB03102, DB03670, DB07298	P18031-215
DB02200–P18031-215	DB03483, DB03557, DB03714, DB06887, DB07651, DB07719, DB08003, DB08549, DB08783	P18031-215
DB00171–P15121-299	DB02338, DB03461	P15121-299
DB00155–P43235-139	DB04276, DB04523, DB07592	P43235-139

4.3. Discussion

For drug discovery and development, it is one of most important things to find protein targets of drugs. Although a large number of approaches have been proposed to detect new targets of drugs in the past decades [5–7,10,11,14,71–78], few have dealt with specific proteins such as post-translationally modified proteins which play key regulating roles in cellular activities [79,80]. Protein SNO is involved in the pathological progression of some diseases, especially neurodegenerative diseases. Therefore, identifying associations between SNO-proteins and small molecules is helpful, especially to develop and discover new drugs treating SNO-mediated diseases. For the first time, we have developed a computational method to predict associations between SNO-proteins and small molecules. This method achieved the expected performances on the experimental dataset. We compared the contributions of various fingerprints of molecules to the recognition of associations between drugs and SNO-proteins. Of all the compared fingerprints, the MACCS obtained the best performances (Figure 3), while the combined fingerprints and the hybridization performed worst and second-worst, respectively (Figure 3). Different fingerprints contributed differently, suggesting specificity concerning protein-binding structure. The reason why MACCS fingerprints perform better than the others is unknown, but it is clear that there would be a certain key substructure associated closely with SNO-proteins. Using this fingerprint would promote the performance of predicting associations between drugs and SNO-proteins. Flooding of the informative fingerprints by a large number of irrelative fingerprints would explain why the combination of all fingerprints performed the worst.

In the areas of bioinformatics, there is a common hypothesis called the GBA principle [66]. The computational analysis of drug–drug similarities and protein–protein similarities supported this hypothesis. This provides an idea of finding an SNO-protein-associated drug or a drug-associated SNO-protein from its analogues. However, there are some exceptions observed in the training set. For example, the compounds phosphoaminophosphonic acid guanylate ester (DB02082) and 7*N*-methyl-8-hydroguanosine-5′-diphosphate (DB01960) have a similarity value of 0.8358, but they were associated with different SNO-proteins. Namely, DB01960 is associated with the proteins eukaryotic translation initiation factor 4E type 3 (Q9DBB5), and DB02082 with Ras-related C3 botulinum toxin substrate 1 (P63000). The binding of the drug to the target might be subject to key substructures. Although highly similar, the small molecules DB01960 and DB02082 might lack common key substructures to bind proteins.

5. Conclusions

SNO-proteins are involved in the pathological processes of some diseases. There is no doubt that the identification of drug–SNO associations is helpful to discover and develop new drugs to treat SNO-mediated disease. We, for the first time, present a machine learning-based computational method to predict associations between drugs and SNO-proteins. The method is simple to implement but quite efficient. We statistically showed that SNO-proteins associated with the same drug would share high similarity, and vice versa. The finding would be useful in detecting drug-associated SNO-proteins and SNO-protein-associated drugs by the similarity search.

Supplementary Materials: The following are available online.

Acknowledgments: This work is supported by the National Natural Science Foundation of China (61672356), by the Provincial Natural Science Foundation of Hunan (2017JJ2239), by the Scientific Research Fund of the Hunan Provincial Science and Technology Department (2014SK2015), and by the aid program for the Science and Technology Innovative Research Team in the Higher Educational Institutions of Hunan Province.

Author Contributions: G.H., J.L., and C.Z. conceived and designed the experiments; G.H. performed the experiments; G.H., J.L., and C.Z. analyzed the data; G.H. contributed reagents/materials/analysis tools; G.H., J.L., and C.Z. wrote the paper

Conflicts of Interest: The authors declare no competing financial interest.

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2016. *CA Cancer J. Clin.* **2016**, *66*, 7–30. [CrossRef] [PubMed]
2. Craig, J. Complex Diseases: Research and Applications. *Nat. Educ.* **2008**, *1*, 184.
3. Burns, A.; Iliffe, S. Alzheimer's disease. *BMJ* **2009**, *338*, b158. [CrossRef] [PubMed]
4. Adams, C.P.; Brantner, V.V. Estimating the cost of new drug development: Is it really $802 million? *Health Aff.* **2006**, *25*, 420–428. [CrossRef] [PubMed]
5. Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S. DINIES: Drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.* **2014**, *42*, W39–W45. [CrossRef] [PubMed]
6. Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232–i240. [CrossRef] [PubMed]
7. Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **2010**, *26*, i246–i254. [CrossRef] [PubMed]
8. Sawada, R.; Kotera, M.; Yamanishi, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. *Mol. Inform.* **2014**, *33*, 719–731. [CrossRef] [PubMed]
9. Chen, X.; Yan, C.C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug-target interaction prediction: Databases, web servers and computational models. *Brief. Bioinform.* **2016**, *17*, 696–712. [CrossRef] [PubMed]
10. Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503. [CrossRef] [PubMed]
11. Chen, H.; Zhang, Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS ONE* **2013**, *8*, e62975. [CrossRef] [PubMed]
12. Chen, X.; Liu, M.X.; Yan, G.Y. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* **2012**, *8*, 1970–1978. [CrossRef] [PubMed]
13. Yildirim, M.A.; Goh, K.I.; Cusick, M.E.; Barabasi, A.L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126. [CrossRef] [PubMed]
14. Huang, G.; Feng, K.; Li, X.; Peng, Y. Large-Scale Prediction of Drug Targets Based on Local and Global Consistency of Chemical-Chemical Networks. *Comb. Chem. High Throughput Screen.* **2016**, *19*, 121–128. [CrossRef] [PubMed]
15. Gao, Y.F.; Chen, L.; Huang, G.H.; Zhang, T.; Feng, K.Y.; Li, H.P.; Jiang, Y. Prediction of drugs target groups based on ChEBI ontology. *BioMed Res. Int.* **2013**, *2013*, 132724. [CrossRef] [PubMed]
16. Wang, Y.C.; Zhang, C.H.; Deng, N.Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.* **2011**, *35*, 353–362. [CrossRef] [PubMed]
17. Xia, Z.; Wu, L.Y.; Zhou, X.; Wong, S.T. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* **2010**, *4* (Suppl. 2), S6. [CrossRef] [PubMed]
18. Vina, D.; Uriarte, E.; Orallo, F.; Gonzalez-Diaz, H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Mol. Pharm.* **2009**, *6*, 825–835. [CrossRef] [PubMed]
19. Mei, J.P.; Kwoh, C.K.; Yang, P.; Li, X.L.; Zheng, J. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics* **2013**, *29*, 238–245. [CrossRef] [PubMed]
20. Eichmann, C.; Tzitzilonis, C.; Nakamura, T.; Kwiatkowski, W.; Maslennikov, I.; Choe, S.; Lipton, S.A.; Riek, R. S-Nitrosylation Induces Structural and Dynamical Changes in a Rhodanese Family Protein. *J. Mol. Biol.* **2016**, *428*, 3737–3751. [CrossRef] [PubMed]
21. Foster, M.W.; Hess, D.T.; Stamler, J.S. Protein S-nitrosylation in health and disease: A current perspective. *Trends Mol. Med.* **2009**, *15*, 391–404. [CrossRef] [PubMed]
22. Kim, J.; Won, J.S.; Singh, A.K.; Sharma, A.K.; Singh, I. STAT3 regulation by S-nitrosylation: Implication for inflammatory disease. *Antioxid. Redox Signal.* **2014**, *20*, 2514–2527. [CrossRef] [PubMed]

23. Nakamura, T.; Tu, S.; Akhtar, M.W.; Sunico, C.R.; Okamoto, S.; Lipton, S.A. Aberrant protein s-nitrosylation in neurodegenerative diseases. *Neuron* **2013**, *78*, 596–614. [CrossRef] [PubMed]
24. Zahid, S.; Khan, R.; Oellerich, M.; Ahmed, N.; Asif, A.R. Differential S-nitrosylation of proteins in Alzheimer's disease. *Neuroscience* **2014**, *256*, 126–136. [CrossRef] [PubMed]
25. Nakamura, T.; Prikhodko, O.A.; Pirie, E.; Nagar, S.; Akhtar, M.W.; Oh, C.K.; McKercher, S.R.; Ambasudhan, R.; Okamoto, S.; Lipton, S.A. Aberrant protein S-nitrosylation contributes to the pathophysiology of neurodegenerative diseases. *Neurobiol. Dis.* **2015**, *84*, 99–108. [CrossRef] [PubMed]
26. Zhao, Q.F.; Yu, J.T.; Tan, L. S-Nitrosylation in Alzheimer's disease. *Mol. Neurobiol.* **2015**, *51*, 268–280. [CrossRef] [PubMed]
27. Wang, Z. Protein S-nitrosylation and cancer. *Cancer Lett.* **2012**, *320*, 123–129. [CrossRef] [PubMed]
28. Ben-Lulu, S.; Ziv, T.; Weisman-Shomer, P.; Benhar, M. Nitrosothiol-Trapping-Based Proteomic Analysis of S-Nitrosylation in Human Lung Carcinoma Cells. *PLoS ONE* **2017**, *12*, e0169862. [CrossRef] [PubMed]
29. Nakamura, T.; Lipton, S.A. Protein S-Nitrosylation as a Therapeutic Target for Neurodegenerative Diseases. *Trends Pharmacol. Sci.* **2016**, *37*, 73–84. [CrossRef] [PubMed]
30. Huang, K.Y.; Su, M.G.; Kao, H.J.; Hsieh, Y.C.; Jhong, J.H.; Cheng, K.H.; Huang, H.D.; Lee, T.Y. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.* **2016**, *44*, D435–D446. [CrossRef] [PubMed]
31. Lu, C.T.; Huang, K.Y.; Su, M.G.; Lee, T.Y.; Bretana, N.A.; Chang, W.C.; Chen, Y.J.; Chen, Y.J.; Huang, H.D. DbPTM 3.0: An informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **2013**, *41*, D295–D305. [CrossRef] [PubMed]
32. Consortium, U. UniProt: A hub for protein information. *Nucleic Acids Res.* **2014**, *43*, D204–D212. [CrossRef] [PubMed]
33. Magrane, M.; Consortium, U. UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011**, *2011*, bar009. [CrossRef] [PubMed]
34. UniProt, C. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142–D148.
35. UniProt, C. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2014**, *42*, D191–D198.
36. UniProt, C. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **2013**, *41*, D43–D47.
37. Breiman, L. Random forests. *MLear* **2001**, *45*, 5–32.
38. Christie, B.D.; Leland, B.A.; Nourse, J.G. Structure searching in chemical databases by direct lookup methods. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 545–547. [CrossRef]
39. Johnson, M.A.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, NY, USA, 1990.
40. McGregor, M.J.; Pallai, P.V. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448. [CrossRef]
41. Hall, L.H.; Kier, L.B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045. [CrossRef]
42. Bolton, E.E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H. Chapter 12—PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Wheeler, R.A., Spellmeyer, D.C., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Volume 4, pp. 217–241.
43. Chen, B.; Wild, D.; Guha, R. PubChem as a source of polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044–2055. [CrossRef] [PubMed]
44. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef] [PubMed]
45. Cereto-Massague, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallve, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63. [CrossRef] [PubMed]
46. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef] [PubMed]
47. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500. [CrossRef] [PubMed]
48. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E.L. Recent developments of the chemistry development kit (CDK)—An open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120. [CrossRef] [PubMed]

49. Dong, J.; Cao, D.-S.; Miao, H.-Y.; Liu, S.; Deng, B.-C.; Yun, Y.-H.; Wang, N.-N.; Lu, A.-P.; Zeng, W.-B.; Chen, A.F. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* **2015**, *7*, 60. [CrossRef] [PubMed]
50. Yu, C.; Deng, M.; Zheng, L.; He, R.L.; Yang, J.; Yau, S.S. DFA7, a new method to distinguish between intron-containing and intronless genes. *PLoS ONE* **2014**, *9*, e101363. [CrossRef] [PubMed]
51. Yu, C.; Deng, M.; Cheng, S.Y.; Yau, S.C.; He, R.L.; Yau, S.S. Protein space: A natural method for realizing the nature of protein universe. *J. Theor. Biol.* **2013**, *318*, 197–204. [CrossRef] [PubMed]
52. Cai, Y.D.; Ricardo, P.W.; Jen, C.H.; Chou, K.C. Application of SVM to predict membrane protein types. *J. Theor. Biol.* **2004**, *226*, 373–376. [CrossRef] [PubMed]
53. Carr, K.; Murray, E.; Armah, E.; He, R.L.; Yau, S.S. A rapid method for characterization of protein relatedness using feature vectors. *PLoS ONE* **2010**, *5*, e9550. [CrossRef] [PubMed]
54. Li, B.Q.; Zhang, Y.C.; Huang, G.H.; Cui, W.R.; Zhang, N.; Cai, Y.D. Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. *PLoS ONE* **2014**, *9*, e86729. [CrossRef] [PubMed]
55. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [CrossRef] [PubMed]
56. Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A.E.; Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* **2018**, *34*, 398–406. [CrossRef] [PubMed]
57. Zhang, N.; Li, B.-Q.; Gao, S.; Ruan, J.-S.; Cai, Y.-D. Computational prediction and analysis of protein γ-carboxylation sites based on a random forest method. *Mol. Biosyst.* **2012**, *8*, 2946–2955. [CrossRef] [PubMed]
58. Hamby, S.E.; Hirst, J.D. Prediction of glycosylation sites using random forests. *BMC Bioinform.* **2008**, *9*, 500. [CrossRef] [PubMed]
59. Ijaz, A. SUMOhunt: Combining Spatial Staging between Lysine and SUMO with Random Forests to Predict SUMOylation. *ISRN Bioinform.* **2013**, *2013*, 671269. [CrossRef] [PubMed]
60. Trost, B.; Kusalik, A. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* **2013**, *29*, 686–694. [CrossRef] [PubMed]
61. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
62. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice-Hall: Egnlewood Cliffs, NJ, USA, 1995; Volume 25.
63. Schwenker, F.; Kestler, H.A.; Palm, G. Three learning phases for radial-basis-function networks. *Neural Netw.* **2001**, *14*, 439–458. [CrossRef]
64. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [CrossRef]
65. Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2007**, *14*, 1–37. [CrossRef]
66. Oliver, S. Proteomics: Guilt-by-association goes global. *Nature* **2000**, *403*, 601–603. [CrossRef]
67. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
68. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204. [CrossRef] [PubMed]
69. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091–D1097. [CrossRef] [PubMed]
70. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; et al. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–D1041. [CrossRef]
71. Alaimo, S.; Pulvirenti, A.; Giugno, R.; Ferro, A. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* **2013**, *29*, 2004–2008. [CrossRef] [PubMed]
72. Gonen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **2012**, *28*, 2304–2310. [CrossRef] [PubMed]
73. He, Z.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE* **2010**, *5*, e9603. [CrossRef] [PubMed]

74. Wang, Y.; Zeng, J. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* **2013**, *29*, i126–i134. [CrossRef] [PubMed]
75. Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266. [CrossRef] [PubMed]
76. Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y. Drug target prediction using adverse event report systems: A pharmacogenomic approach. *Bioinformatics* **2012**, *28*, i611–i618. [CrossRef] [PubMed]
77. Bleakley, K.; Yamanishi, Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* **2009**, *25*, 2397–2403. [CrossRef] [PubMed]
78. Dunkel, M.; Gunther, S.; Ahmed, J.; Wittig, B.; Preissner, R. SuperPred: Drug classification and target prediction. *Nucleic Acids Res.* **2008**, *36*, W55–W59. [CrossRef] [PubMed]
79. Jia, C.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* **2018**. [CrossRef] [PubMed]
80. Wei, L.; Xing, P.; Shi, G.; Ji, Z.L.; Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, 1. [CrossRef] [PubMed]

Sample Availability: Not Available.