Predicting Therapeutic Targets with Integration of Heterogeneous Data Sources

Yan-Fen Dai[1,2], Yin-Ying Wang[1,3], and Xing-Ming Zhao[4,⋆]

[1] Institute of Systems Biology, Shanghai University, Shanghai 200444, China
[2] Department of Mathematics, Shanghai University, Shanghai 200444, China
[3] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[4] School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
xm_zhao@tongji.edu.cn

Abstract. Drug target is of great importance for designing new drugs and understanding the molecular mechanism of drug actions. In general, a drug may bind to multiple proteins, some of which are not related to disease-treatment or even lead to side effects. Therefore, it is necessary to discriminate the effect-mediating drug targets, i.e. therapeutic targets, from other proteins. Although a lot of computational approaches have been developed to predict drug targets and achieve partial success, few attention has been paid to predict therapeutic targets. In this work, we present a new framework to predict drug therapeutic targets based on the integration of heterogeneous data sources. In particular, we develop an ensemble classifier, PTEC (Predicting Therapeutic targets with Ensemble Classifier), that can effeciently integrate both drug and protein properties described from distinct perspectives, thereby improving prediction accuracy. The results on benchmark datasets demonstrate that our approach outperforms other popular approaches significantly, implying the effectiveness of our proposed approach. Furthermore, the results indicate that the integration of different data sources can not only improve the coverage of predicted targets but also the prediction precision. In other words, distinct data sources indeed complement with each other, and the integration of these heterogeneous data sources can improve the prediction accuracy.

1 Introduction

Drug target identification is one of the most important steps in drug development, and is the key to understand how the desirable therapeutic effects are accomplished when the proteins are targeted by drugs [1,2]. Unfortunately, the targets of a lot of drugs are incomplete or even unknown, which hampers the discovery of new drugs. Recently, a number of computational approaches have been proposed to predict drug targets. For example, assuming similar drugs bind

⋆ Corresponding author.

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 149–158, 2013.

to similar pockets on the protein surfaces, molecular docking approaches have been widely used to identify those compounds that can bind to known target proteins by investigating the chemical similarity between candidate ligands with known drugs [3]. With the knowledge that drugs with similar therapeutic effects generally target same proteins, drug therapy information has been used to predict drug targets [4]. Observing that drugs with similar side effects tend to target common proteins, Campillos *et al* proposed a novel approach to predict drug targets based on side effect similarity [5]. Considering that protein function is determined by its component domains while ligands generally bind to proteins to exert their function [6], Wang *et al* proposed a novel statistical approach to predict drug targets based on the derived interactions between drugs and protein domains [7]. To further improve prediction accuracy, different kinds of data sources have been integrated to predict compound-protein interactions. For example, Yamanishi *et al* have combined chemical structure and genomic sequence information to predict drug-protein interactions [8], and they later further took into account the pharmacological information to improve prediction accuracy [9].

With the knowledge about drug-protein interactions becoming more comprehensive, the amount of compound-protein interactions deposited in public databases, e.g. DrugBank [10] and STITCH [11], increases accordingly. Most recently, it is found that actually 96% of approved drugs have known targets [12]. However, a large number of these drug-protein interactions are found to be either irrelevant to disease-treatment or related to side effects [13]. In general, a compound may bind to multiple proteins, among which some proteins are off-targets that may lead to severe undesirable adverse effects. That is, druggable proteins are not necessarily main effect-mediating targets, i.e. therapeutic targets, that play critical and preferably unsubstitutable roles when treating disease [14]. Therefore, it is necessary to identify those therapeutic targets, and discriminate them from therapeutically irrelevant or side effect related ones. The therapeutic targets can help design drugs with expected efficacy. Although experimental techniques, such as high-throughout screening with bioassays, can be used to detect drug-protein interactions, it is highly expensive and time-consuming to identify the effect-mediating targets from the large pool of proteins within the human genome. Despite the partial success achieved by above mentioned computational approaches, few attention has been paid to predict therapeutic targets in the bioinformatics community possibly due to the scarceness of therapeutic target information.

In this paper, we present a novel framework to predict the therapeutic targets for known drugs based on integration of heterogeneous data sources. To this end, we investigate various properties of both drugs and proteins, including chemical structure and therapy information for drugs while primary structure and functional annotations for proteins. In particular, we develop a novel approach to integrate these heterogeneous data for both drugs and proteins with an ensemble classifier, PTEC (Predicting Therapeutic targets with Ensemble Classifier). The integration of different data sources can not only improve prediction

coverage but also accuracy [15]. That is, distinct data sources can complement with each other so that better results are expected based on the integration of these heterogeneous data sources. The results on gold standard datasets demonstrate that our proposed method outperforms other popular approaches significantly, implying the effectiveness of our proposed approach.

The rest of this paper is organized as following. Section 2 presents the materials used in this work and our proposed methods; Section 3 presents the experimental results; Finally, conclusions are drawn in Section 4.

2 Materials and Methods

2.1 Data Sources

In this work, 406 therapeutic targets for known drugs were retrieved from [12], which were curated from the drug-protein interactions from the DrugBank database [10]. We also downloaded other human drug target proteins and drug therapy information from DrugBank database (version 3.0). The drug therapy information described as therapeutic categories in Anatomic Therapeutic Chemical (ATC) classification system was considered here. The chemical structure information for drugs was obtained from PubChem [16]. As a result, 708 drugs with both chemical structure and therapy information available were kept for further analysis, which leads to 1726 interactions between drugs and their corresponding therapeutic targets.

The amino acid sequences of human proteins were obtained from the Uniprot database [17]. The functional annotations for these proteins were extracted from the Gene Ontology (GO) database [18], where all three functional categories were considered, including cellular component, molecular function and biological process. The protein associated pathway information was retrieved from KEGG database [19]. Furthermore, the expression profiles of protein coding genes generated for 36 normal human tissues were obtained from [20].

2.2 Drug Similarity

With chemical structure and therapy information available for drugs, we can define the similarity between two drugs. The chemical similarity between a pair of drugs was calculated as the two-dimensional Tanimoto score based on their fingerprints with the help of Chemistry Development Kit (CDK) [21], which is defined as following.

$$C_s(d, d^{'}) = \frac{\sum_i (d_i \wedge d_i^{'})}{\sum_j (d_j \vee d_j^{'})} \tag{1}$$

where $C_s(d, d^{'})$ represents the similarity score of two drugs d and $d^{'}$, d_i is the ith bit in the fingerprint of drug d, and $\wedge$ and $\vee$ respectively denotes bitwise $'and'$ and $'or'$ operators.

In the Anatomic Therapeutic Chemical (ATC) classification system, each drug can be described in 5 hierarchical levels and is classified into different therapeutic

groups according to the organ it acts on and its chemical characteristics. In this work, the therapeutic similarity between two drugs was defined as their longest matched prefix between their corresponding ATC codes as described previously [4].

$$T(d, d^{'}) = \max_{(d_i, d_j^{'})} \frac{2 * \log(Pr(pre(d_i, d_j^{'})))}{\log(Pr(d_i)) + \log(Pr(d_j^{'}))} \tag{2}$$

where $T(d, d^{'})$ denotes the therapeutic similarity between drugs d and $d^{'}$, d_i denotes the ith ATC category for drug d considering each drug may be grouped into different categories, $pre(i, j)$ denotes the longest matched prefix between the ATC codes d_i and $d_j^{'}$, $Pr(d_i)$ denotes the probability of the ATC category d_i occurs in drugs, and $Pr(pre(d_i, d_j^{'}))$ denotes the probability of the common prefix between the two ATC categories d_i and $d_j^{'}$ occurs in drugs.

2.3 Protein Similarity

The most straightforward way to measure the similarity between two proteins is to compare their primary structure identity. In this work, the sequence similarity $S_s(p, p^{'})$ between two proteins $(p, p^{'})$ is defined as the normalized Smith-Waterman alignment score as described as following.

$$S_s(p, p^{'}) = \frac{SS(p, p^{'})}{\sqrt{SS(p, p)SS(p^{'}, p^{'})}} \tag{3}$$

where $SS(.,.)$ denotes the original Smith-Waterman alignment score [22].

The pathways associated with drug target proteins can tell the molecular context in which the proteins exert their function, and therefore help to understand the mechanism of actions of drugs. With pathway annotation for proteins available, the pathway similarity $S_p(p, p^{'})$ between two proteins can be defined as below.

$$S_p(p, p^{'}) = \frac{\left| S(p) \cap S(p^{'}) \right|}{|S(p) \cup S(p^{'})|} \tag{4}$$

where $S(p)$ and $S(p^{'})$ respectively denotes the set of pathways in which protein p and $p^{'}$ are located.

Furthermore, with the functional annotations extracted from GO database, the functional similarity $S_g(p, p^{'})$ between two proteins p and $p^{'}$ is defined as the Jaccard index.

$$S_g(p, p^{'}) = \frac{\sum_{k=1}^{3} \left| t_k(p) \cap t_k(p^{'}) \right|}{\sum_{k=1}^{3} |t_k(p) \cup t_k(p^{'})|} \tag{5}$$

where $t_k(p)$ is the set of GO terms associated with protein p with respect to functional category k, $k = 1, 2, 3$ denotes each of the three functional categories in GO database, i.e. Molecular Function, Biological Process, and Cellular Component.

In addition, the expression similarity between two genes coding a pair of proteins was defined as coexpression correlation based on the gene expression profiles of 36 normal human tissues from [20] as below.

$$S_t(p, p^{'})) = \frac{\sum_{k=1}^{n}(p(k)-\overline{p})(p^{'}(k)-\overline{p^{'}})}{\sum_{k=1}^{n}\sqrt{(p(k)-\overline{p})^2(p^{'}(k)-\overline{p^{'}})^2}} \tag{6}$$

where $S_t(p, p^{'})$ is the correlation coefficient between the genes coding proteins p and $p^{'}$, n is the number of samples, and $\overline{p}$ is the mean of expression profile of protein p.

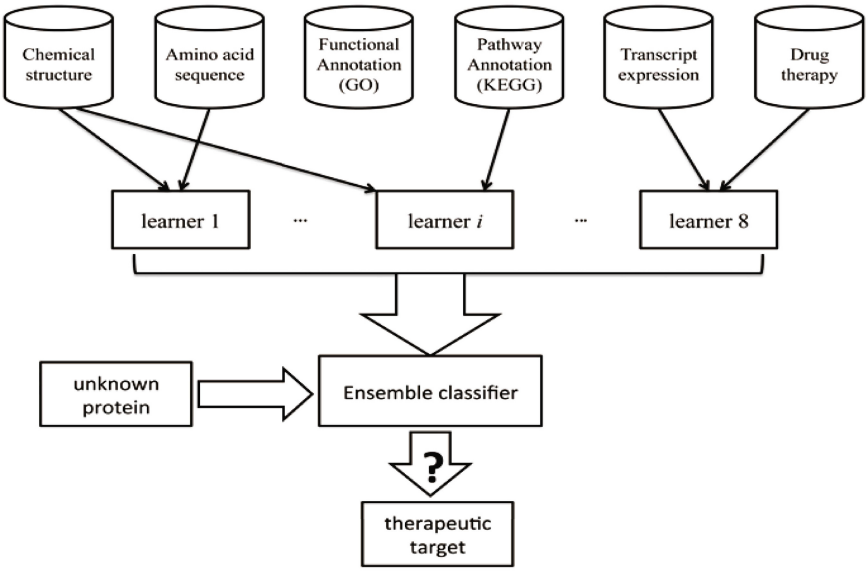

Fig. 1. The flowchart of predicting therapeutic targets based on the integration of heterogeneous data sources

2.4 Therapeutic Target Prediction

With the drug similarity described above, we assume that drugs with similar characteristics will target same proteins. Similarly, the proteins with similar properties will be bound by same drugs. With this in mind, we can construct a learner based on known drug-protein interactions. In this work, a drug-protein pair (d_i, p_j) can be represented as a feature vector (Fd_i, Fp_j), where each element in Fd_i represents the similarity between drug d_i and all the drugs while each element in Fp_j represents the similarity between protein p_i and all the proteins. For example, for the combination of chemical structure and protein

sequence, the elements in Fd_i denotes the chemical similarity between drug d_i and the rest drugs while the elements in Fp_j denotes the sequence similarity between protein p_j and the other proteins. After the feature extraction step, a classifier will be subsequently trained for each combination of drug and protein properties, e.g. drug therapy and protein sequence. In this way, we can have 8 different combinations between distinct drug and protein properties, thereby leading to 8 classifiers. Instead of selecting the best-performing classifier from the eight ones, we proposed to construct an ensemble classifier, PTEC (Predicting Therapeutic targets with Ensemble Classifier), to integrate these distinct learners in a weighted way (see Fig 1). The ensemble classifier was adopted here since it has been found to outperform individual ones and is more robust [23]. In particular, we first evaluated each classifier on a benchmark dataset, and used their accuracy as their corresponding weights to construct the ensemble classifier as following.

$$Enc_{res} = \sum_{i=1}^{8} W_i \cdot L_i \tag{7}$$

where Enc_{res} is the predicted results by the ensemble classifier, W_i is the weight for learner ith, and L_i is the output of learner ith. Here the weight for each learner is set to the area under the curve (AUC) score of a reciever operating characteristic (ROC) curve it obtained on the training set. Therefore, for a given unknown protein, we can use the Ensemble classifier to predict whether it is a therapeutic target. The simple but effective k-nearest neighbor algorithm (k-NN) was used as the learner in this work.

3 Results and Discussion

With the known interactions between drugs and their corresponding therapeutic targets as positive set, we build a negative set consists of drug-protein interactions from DrugBank except those from the positive set for the drugs involved in the positive set. As a result, 1094 drug-protein interactions were obtained as negative set. Note that all the drug-protein interactions in the negative set are real interactions as reported in DrugBank.

To evaluate the predictive power of different classifiers, one fifth of the samples were used as the test set while the rest were used as the training set. Firstly, we evaluated the eight single classifiers based on the training set with 10-fold cross-validation. Table 1 summarizes the results obtained by distinct classifiers. From the results, we can see that these eight classifiers perform comparably well with no one single classifier performs always best. For example, the classifier trained with therapy information and gene expression achieves the highest true positive rate, while the one trained on protein sequence performs best with respect to false negative rate. With the AUC scores obtained by the eight classifiers on the training set as their corresponding weights, we integrated the eight classifiers into an ensemble classifier PTEC, which achieves the highest true positive rate and the best overall result with an AUC score of 0.71 (see Table 1). The ensemble classifier

Table 1. Performance of distinct classifiers, where the results were obtained with 10-fold cross-validation on the training set

	C_{cs}	C_{cp}	C_{ct}	C_{cg}	C_{As}	C_{Ap}	C_{At}	C_{Ag}	**PTEC**
TPR	0.77	0.76	0.80	0.76	0.77	0.79	0.80	0.79	**0.81**
TPR_*std*	0.02	0.01	0.02	0.02	0.03	0.02	0.01	0.01	0.01
FPR	0.37	0.41	0.46	0.36	**0.37**	0.43	0.46	0.43	0.39
FPR_*std*	0.02	0.01	0.03	0.02	0.02	0.01	0.02	0.02	0.01
AUC	0.70	0.66	0.68	0.70	0.68	0.67	0.70	0.70	**0.71**
AUC_*std*	0.02	0.01	0.02	0.02	0.01	0.01	0.02	0.01	0.01

C_{cs} - classifier trained on chemical structure and protein sequence; C_{cp} - classifier trained on chemical structure and protein pathway; C_{ct} - classifier trained on chemical structure and transcriptional expression; C_{cg} - classifier trained on chemical structure and protein GO annotation; C_{As} - classifier trained on drug therapy information and protein sequence; C_{Ap} - classifier trained on therapy information and protein pathway; C_{At} - classifier trained on therapy information and transcriptional expression; C_{Ag} - classifier trained on therapy information and protein GO annotation;
TPR - true positive rate;
TPR_*std* - standard deviation of true positive rate;
FPR - false positive rate;
FPR_*std* - standard deviation of false positive rate;
AUC - Area under ROC curve;
AUC_*std* - standard deviation of AUC.

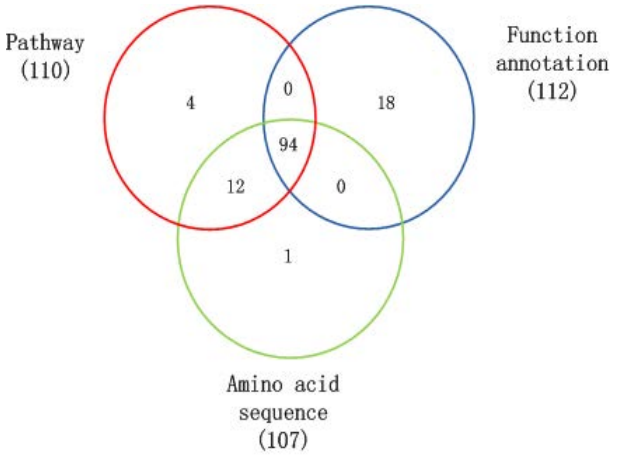

Fig. 2. The Venn diagram about the number of drug-protein interactions successfully predicted by the combination between drug therapy and three protein properties

was adopted here since it can improve prediction coverage considering that the annotations for proteins are incomplete. For example, looking into the drug-protein interactions predicted by different classifiers, Fig. 2 shows the Venn diagram about the number of drug-target pairs successfully predicted by the combination of drug therapy with protein sequence, pathway annotation and functional annotation respectively. It can be seen that among the 138 drug-protein interactions, the three

Table 2. Performance of distinct classifiers on the test set

	C_{cs}	C_{cp}	C_{ct}	C_{cg}	C_{As}	C_{Ap}	C_{At}	C_{Ag}	**PTEC**
TPR	0.70	0.78	0.79	0.80	0.76	0.80	0.82	0.82	**0.83**
FPR	0.26	0.26	0.36	0.29	0.32	0.33	0.38	0.34	**0.17**
AUC	0.76	0.75	0.71	0.75	0.73	0.74	0.72	0.74	**0.83**

classifiers get consistent results on most of their predictions 68.12% (94/138), while the integration of these different data sources can enlarge the number of predicted therapeutic targets significantly. In other words, distinct data sources complement with each other and the integration of them can improve both prediction accuracy and coverage.

To further evaluate the predictive power of our proposed PTEC, we applied it to predict therapeutic targets on the hold-out test set. Moreover, we compared our results with those eight single classifiers. Table 3 shows the performance of distinct classifiers on the test set. The results demonstrate that our proposed ensemble classifier significantly outperforms others with an AUC score of 0.83 and the highest true positive rate, indicating the effectiveness and robustness of our proposed ensemble classifier.

In addition, we compared our proposed method with a popular approach, namely nearest profile (NP), which predicts drug targets based on a bipartite graph. Figure 3 gives the results obtained by both PTEC and NP, where the

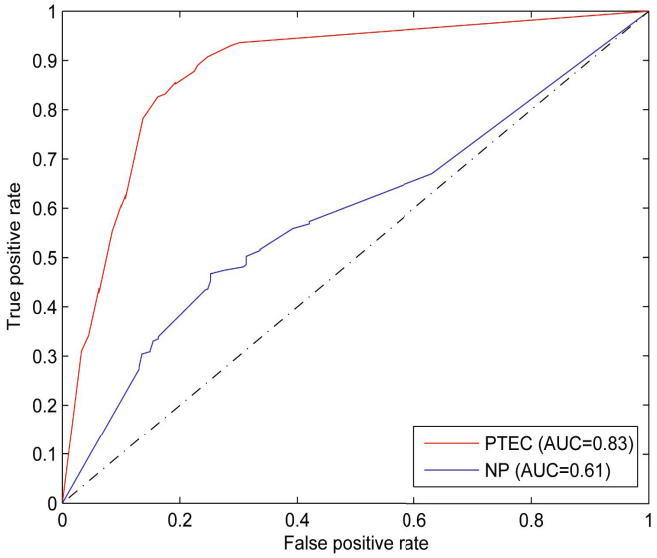

Fig. 3. The performance of PTEC and the nearest profile(NP) method

results by PTEC are based on the test set while those by NP are based on the whole dataset. From the results, we can clearly see that PTEC is really effective to predict therapeutic targets, and is able to separate therapeutic targets from other irrelevant ones. The good performance of PTEC confirm again that the integration of different data sources indeed can improve prediction accuracy and also the predictive power of our proposed approach.

In our predictions, some of them are not found in the positive dataset, which does not necessarily mean they are false positives. For example, we predict protein AchE that is involved in lipid transportation and metabolism as the therapeutic target of drug Physostol, a cholinesterase inhibitor that can be applied topically to the conjunctiva. In the positive set, AchE is not the therapeutic target of Physostol, whereas we found that AchE is reported as the therapeutic target of Physostol in the Therapeutic Target Database (TTD)[24]. The drug Metubine iodide is a benzylisoquinolinium competitive nondepolarizing neuromuscular blocking agent, which was predicted to bind to CHRNA2 by our proposed PTEC, and this interaction is also verified in TTD. The verification of our prediction results by other public databases demonstrates the predictive power of our proposed method.

4 Concluding Remarks

Therapeutic target is the key to design the drugs with expected efficiency and understand how the drugs work. In this paper, we present a new framework to predict drug therapeutic targets by integrating heterogeneous data sources for both drugs and proteins. Specifically, we proposed a novel ensemble classifier to integrate the learners trained on distinct data sources. The results on benchmark dataset demonstrate the effectiveness and robustness of our proposed approach.

Acknowledgement. This work was partly supported by the National Natural Science Foundation of China (91130032, 61103075), Innovation Program of Shanghai Municipal Education Commission (13ZZ072) and Innovation Program of Shanghai University(SHUCX120115).

References

1. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, 1119–1126 (2007)
2. Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., et al.: Analysis of multiple compound-protein interactions reveals novel bioactive molecules. Mol. Syst. Biol. 7, 472 (2011)
3. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., et al.: Predicting new molecular targets for known drugs. Nature 462, 175–181 (2009)
4. Zhao, S., Li, S.: Network-based relating pharmacological and genomic spaces for drug target identification. PLoS ONE 5, e11764 (2010)

5. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P., et al.: Drug target identification using side-effect similarity. Science 321, 263–266 (2008)
6. Zhao, X.M., Chen, L., Aihara, K.: A discriminative approach for identifying domain-domain interactions from protein-protein interactions. Proteins 78, 1243–1253 (2010)
7. Wang, Y.Y., Nacher, J.C., Zhao, X.M.: Predicting drug targets based on protein domains. Mol. Biosyst. 8, 1528–1534 (2012)
8. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatic 24, i232–i240 (2008)
9. Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatic 26, i246–i254 (2010)
10. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic. Acids. Res. 34, D668–D672 (2006)
11. Kuhn, M., Szklarczyk, D., Franceschini, A., et al.: STITCH 3: zooming in on protein-chemical interactions. Nucleic. Acids. Res. 40, D876–D880 (2012)
12. Gregori-Puigjane, E., Setola, V., Hert, J., Crews, B.A., Irwin, J.J., et al.: Identifying mechanism-of-action targets for drugs and probes. Proc. Natl. Acad. Sci. U S A 109, 11178–11183 (2012)
13. Rask-Andersen, M., Almen, M.S., Schioth, H.: Trends in the exploitation of novel drug targets. Nat. Rev. Drug. Discov. 10, 579–590 (2011)
14. Hopkins, A.L., Groom, C.R.: The druggable genome. Nat. Rev. Drug. Discov. 1, 727–730 (2002)
15. Zhao, X.M., Iskar, M., Zeller, G., Kuhn, M., van Noort, V., Bork, P.: Prediction of drug combinations by integrating molecular and pharmacological data. PLoS. Comput. Biol. 7, e1002323 (2011)
16. Wang, Y.L., Xiao, J.W., Suzek, T.O., et al.: PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic. Acids. Res. 37, W623–W633 (2008)
17. Apaweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., et al.: UniProt: the Universal Protein knowledgebase. Nucleic. Acids. Res. 32, D115–D119 (2004)
18. Michael, A., Catherine, A.B., Judith, A.B., David, B., Heather, B., et al.: Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000)
19. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic. Acids. Res. 28, 27–30 (2000)
20. Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M., Aburatani, H.: Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86, 127–141 (2005)
21. 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. 43, 493–500 (2003)
22. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197 (1981)
23. Zhao, X.M., Li, X., Chen, L., Aihara, K.: Protein classification with imbalanced data. Proteins 70, 1125–1132 (2008)
24. Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., et al.: Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic. Acids. Res. 40, D1128–D1136 (2012)