Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, 11, 2096

Protein localization vector propagation: a method for improving the accuracy of drug repositioning

Yunku Yeu,[a] Youngmi Yoon[b] and Sanghyun Park*[a]

Received 30th April 2015,
Accepted 15th May 2015

DOI: 10.1039/c5mb00306g

www.rsc.org/molecularbiosystems

Identifying alternative indications for known drugs is important for the pharmaceutical industry. Many computational methods have been proposed for predicting unknown associations between drugs and target proteins associated with diseases. To produce better prediction, researchers should not only develop accurate algorithms but identify good features that reflect intracellular systems. In this paper, we proposed a novel method for exploiting protein localization. We generated localization vectors (LVs) from protein localization and propagated LVs through a protein interaction network to increase the coverage of the localization information. The LVs showed distinct patterns among targets of known drugs as well as independent characteristics compared to existing features. Based on the experimental results, we determined that including LVs improves cross-validation accuracy and, produces better novel predictions with real and independent clinical trial data. Moreover, the propagation of LVs showed a positive result that it can help in increasing the coverage of the prediction results.

1. Introduction

Since the mid-1990s, productivity of pharmaceutical research has been declining. According to an analysis of pharmaceutical research and development in 2000–2008,[1] the average success rate of new molecular entities was 2.01%; moreover, an average of 13.9 years was required for clinical development.

Drug repositioning is an alternative methodology for drug discovery. Researchers strive to find unknown indications or targets for approved or discarded drugs. Successful drug repositioning reduces the time and cost of early development; furthermore it enables drug development to begin directly from testing and clinical trial stages.

To improve the effectiveness of the drug repositioning, *in silico* computational drug repositioning approaches have been adopted for targeting most promising candidates.[2] The approaches are also useful for rare and neglected diseases which have small patient populations. These diseases are hard to be profitable if the cost of drug research and development is large.[3,4]

Computational drug repositioning methods are based on an assumption that the similarities and shared properties between biological entities can help identify their new mechanism of action (MOA).[5] For example, if disease P is an indication of drug D_1, and two drugs D_1 and D_2 have similar characteristics, we can infer that disease P can be a candidate indication for drug D_2. In another view, if D_1 targets a protein T, we can find other candidate target proteins that have characteristics similar to the protein T. These predictions can produce more accurate results when sufficient evidential data and knowledge are included. The available data types for drug repositioning include genome sequences, gene expressions, protein abundance, chemical properties, protein–protein interactions (PPI), network structures, pathways, literature, *etc.*[6]

Numerous computational methods have been proposed for effective drug repositioning. The connectivity map (cmap)[7] exploits differential gene expression patterns as features for recovering connections among drugs, genes, and diseases. Wu *et al.* constructed a network using heterogeneous data and clustered the network to find new drug–disease interactions in the clusters.[8] Cheng *et al.* proposed and evaluated three inference methods that are drug-based, target-based, and network-based.[9] The network-based method showed superior performance. Zhao and Li identified drug–gene–diseases co-modules using a Bayesian partition method; they then inferred new associations in the modules.[10]

A possible approach to improving the effectiveness of the computational methods is to incorporate new features. Jin *et al.* suggested a cancer-related signaling network motif[11] and it was exploited in their drug repositioning study.[12] Sanseau *et al.* suggested GWAS information as a new feature for drug repositioning[13] and Wang *et al.* supported the GWAS through a computational analysis using the pathophysiological information.[14] PREDICT[15] combines multiple features, such as chemical structures and side effects of drugs, sequences of proteins, distances on PPI networks and GO term similarities.

[a] *Department of Computer Science, Yonsei Univertity, 50 Yonsei-Ro, SeoDaeMun-Gu, Seoul, 120-749, Republic of Korea. E-mail: sanghyun@cs.yonsei.ac.kr*
[b] *Department of Computer Engineering, Gachon University, Seongnamdaero 1342, Sujeong-gu, Seongnam-si, Gyeonggi-do, 461-701, Republic of Korea*

Protein localization represents subcellular location(s) where the proteins perform their role. Many proteins have specific localizations, and the protein functions are closely related to their subcellular locus.[16] The localization information was exploited for drug repositioning studies,[17–19] inference of disease comorbidity[20] and protein function predictions.[21] However, to the best of our knowledge, the protein localization has not been exploited effectively in current drug repositioning studies, but has been used simply as a filter that selects membrane and cytoplasmic proteins.[17–19] Although these proteins are good candidates for drug targeting, they may miss many possible candidates which exist in other locations. The GO term (subcellular localization) is a good representation of the protein localization information. However, calculating semantic similarities between GO terms from two proteins needs a very large computation, therefore, it is hard to exploit the GO term in a large size drug repositioning problem.

In this study, we propose a new method for exploiting protein localization as a feature for the drug repositioning research. The localization of a protein was represented as a localization vector (LV). Then the values of the vectors were propagated through the PPI network in order to increase of the localization information. We then examined the characteristics of the propagated protein localization. The propagated LV showed more similar patterns between known target proteins than between random proteins. In addition, we validated the effectiveness of our approach through a test for predicting known drug–target associations. Including LVs in logistic regression classifiers improved the cross-validation accuracy by more than 5% (0.76 → 0.8135). In an independent prediction experiment using a clinical trial database, logistic regression classifiers with LVs predicted nearly two times (195%) as many associations as the prediction result without LVs.

2. Results and discussion

2.1. Data sources

In this study, we used heterogeneous data sources for drugs and target proteins. Drug identifiers and chemical structures expressed in simplified molecular-input line-entry system (SMILES) were extracted from DrugBank.[22] The side effects of drugs were downloaded from SIDER2.[23] Protein data, including identifiers and subcellular locations, were downloaded from Uniprot.[24] Finally, the human-only PPI data were collected from I2D[25] and BioGRID.[26] A summary of data sources is provided in Table 1.

Table 1 Summary of collected data

Data source	Collected data
DrugBank	Drug identifiers, SMILES, target proteins of 7677 drugs
SIDER2	32 140 side effects
Uniprot	Protein identifiers, subcellular locations, amino acid sequences of 136 871 proteins
I2D, BioGrid	231 790 protein–protein interactions

2.2. Localization vectors

At first, we defined the LVs, which represent subcellular localizations of a protein. A LV is a vector with ten attributes, such that each attribute value represents probabilities that the protein will perform its functions at the corresponding subcellular locations. The ten subcellular locations were selected by Park *et al.*[20] including the cytosol, endoplasmic reticulum (ER), extra-cellular, Golgi, peroxisomes, mitochondria, nuclei, lysosomes, plasma membranes, and other locations.

To construct initial LVs, we parsed the Uniprot data and set an LV attribute of a protein to 1 if the protein had a corresponding subcellular location in the Uniprot data. Otherwise, the LV attribute was set to 0. After constructing initial LVs, they were propagated through the PPI network based on an assumption that physically interacting proteins have similar subcellular localizations. (For further details, see Section 4.1.) Through the propagation, the coverage of the localization information could be increased.

The similarity between two LVs was measured using the cosine similarity (CS). CS is an appropriate measure for LVs because it yields the minimum similarity between two LVs that have exclusive (non-overlapping) subcellular locations. These two LVs are orthogonal in the vector space and their CS is 0. A higher LV similarity indicates that the proteins perform their functions in similar subcellular locations.

2.3. Distribution of LVs

In this section, we observed characteristics of the propagated LVs. A distribution of sorted LVs is described in Fig. 1. In the figure, a LV is represented as a horizontal line with ten parts; the brightness of a part represents the value of an LV attribute. For example, the highest part of the figure represents a LV (0,0,0,0,0,0,0,0,1,0) and about 15% of LVs have this pattern. Many LVs existing in the known database used in our study showed simple patterns composed of 0s and 1s (black and white only), while proteins with propagated localizations showed more varied distributions (grey cells in Fig. 1). Note that the information in the grey area of Fig. 1 was unavailable without the propagation.

2.4. Target protein pairs of the same drug *vs.* all pairs

From this section to Section 2.6, we will describe a characteristic of the LV that discriminate the drug–target relationships. First, we examined LV similarities between protein pairs targeted by the same drug. Then the similarities were compared to LV similarities between random protein pairs. A traditional approach for drug repositioning is to find new targets of existing drugs.[11] The effectiveness of a machine learning feature can be indirectly measured by testing its separation ability for known positive and negative samples. We therefore investigated whether the LV is capable of distinguishing the target protein pairs and random protein pairs. For this test, two sets of protein pairs were collected. The first set was generated from multi-target drugs. For each drug that had more than two target proteins, we added all possible pairs within the targets to the first set. The second set, a control group, included all

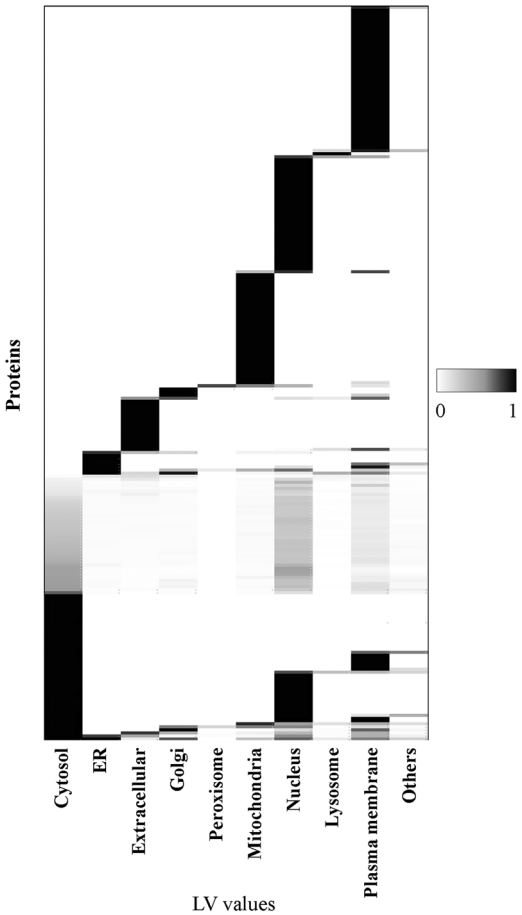

Fig. 1 Distribution of localization vectors. Proteins that have their subcellular information in Uniprot had discrete values (black/white), while proteins that received localization values from their neighbours had continuous values (gray).

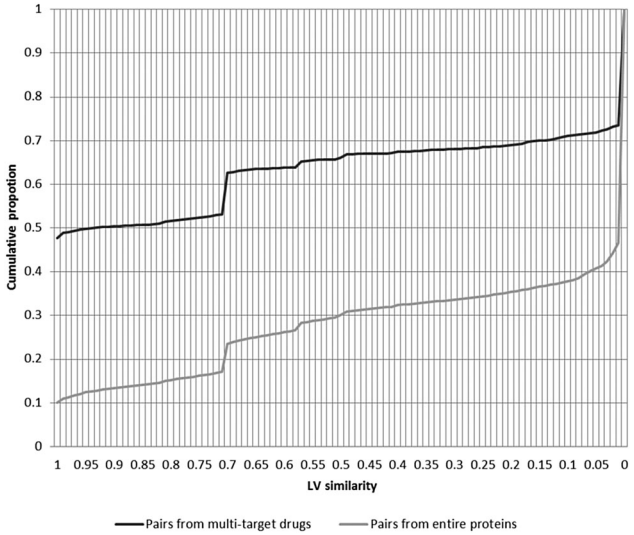

Fig. 2 Cumulative curves for LV similarity distributions from two sets of pairs. Protein pairs extracted from multi-target drugs generated higher LV similarities.

possible pairs of proteins. We then calculated the LV similarity for each pair in the first and second sets. If the distributions of the LV similarities for the two sets were sufficiently different, the LV could serve as an evidence for finding targets of drugs.

Fig. 2 shows two accumulation curves of LV similarities calculated from the two sets of pairs. The similarity values were sorted in decreasing order and accumulated; the higher curve represents the larger composition of high similarity values in the set. For the curve of drug targets (dark line), nearly 50% of protein pairs had greater LV similarities than 0.65, whereas only 25% of all possible protein pairs had LV similarity values greater than 0.65. The AUC (area under curve) of the two curves were 0.566 and 0.278, respectively (2.03 times larger). The average similarities of the two sets of pairs were 0.5613 and 0.4181.

2.5. Comparisons to other features

As discussed in the previous section, LVs demonstrated discriminative characteristics for drug target proteins. However, are not good features if they are dependent on other existing features. In particular, because LVs are propagated from the PPI network, closed pairs in the PPI network may have similar LV patterns. In this section, we compare LVs to existing features for drug repositioning. The existing features include the SMILES similarities of drugs, side effect similarities of drugs, amino acid sequence similarities of proteins, distances in the PPI network, and GO term similarities. To investigate the dependency between the features, we calculated correlations between each feature and the LV similarities.

When we compared LVs with the protein features (sequences, distances in the PPI network, GO terms), we chose protein pairs that have both LVs and the compared features. Then we calculated the similarities and obtained two lists of similarities from the two features. The correlation coefficient was calculated from the lists. For the drug features, such as SMILES and side effects, similarities for these features were generated from a drug pair. Protein feature similarities for the drug pair were generated from the best pair of proteins targeted by the drug pair. We calculated SMILES, side effects, LV similarities and distances in the PPI network for all drug–protein pairs we included. For GO terms similarities, we used the R package[27] and GOSemSim.[28] GOSemSim measures semantic similarities for sets of GO terms. However, it suffers from heavy computational costs and immense time consumption. Thus, we sampled 100 random protein pairs 100 times (a total of 10 000 protein pairs) and applied Resnik's method[29] in three categories: the biological process (BP), molecular function (MF), and cellular component (CC).

Table 2 shows the correlation between the LV and the existing features. The best similar one to the LV is the CC in GO terms. It is a rational result because the LV and the CC represent fundamentally the same information. However, calculating semantic similarities between sets of GO terms is a very complex approach. On the other hand, calculating similarities between LVs is much easier. This is one of the

Table 2 Correlations between LVs and existing features[a]

Existing features	Similarity measures	Correlation with LV similarity
SMILES	Jaccard coefficient	0.0465
Side effect	Jaccard coefficient	0.0733
Amino acid sequence	Smith-Waterman algorithm[30]	−0.0124
PPI distance	Perlman *et al.*[31]	0.1593
GO terms – BP	Resnik *et al.*[29]	0.2947
GO terms – MF		0.2450
GO terms – CC		0.3518

[a] BP: biological process, MF: molecular function, CC: cellular component.

strong points of LVs considering a large number of drug–protein combinations.

In particular, the LV showed a weak correlation with the PPI distance even though the LV was propagated through the PPI network. This is due to the very simple pattern of the PPI distance in a dense network. In our PPI network, the maximum distance in the network was 7 and the PPI distance feature showed only 7 kinds of patterns. The LV shows more various patterns compared to the PPI distance. According to the analysis, we can conclude that the LV is an independent feature for drug repositioning.

2.6. Effects on cross-validation accuracy

In this section, we describe a cross-validation (CV) test that we performed to determine the effectiveness of the LV as a feature for a prediction algorithm. If the LV had a positive effect for the prediction of drug–target associations, the CV accuracy would be improved when the LV was included compared to when it was excluded. We exploited a prediction pipeline proposed by Gottlieb *et al.*[15] SMILES and side effects were included as drug features, and amino acid sequences, PPI distances, and LVs were included as protein features. The GO term similarity was excluded on account of its high computational cost. For the prediction, a logistic regression classifier implemented in Weka[32] was selected. (The detailed CV method is described in Section 4.2.) The gold-standard set from DrugBank was divided into ten subsets. Each subset was considered the positive samples; the same number of random drug–target associations was generated as negative samples. The remaining subsets were regarded as known associations for scoring the positive/negative samples. In each subset, a ten-fold CV was applied to the samples, and accuracy (AUC) of the CV was averaged. The results are presented in Table 3. When all five features were included, the average accuracy of the CV was improved compared to the control case, the dataset without the LV (1st row and 2nd row). When we compared the LV and PPI distance, including the LV yielding a better accuracy (2nd row and 3rd row).

When we compared effects of individual protein features on the prediction, the amino acid sequence was the best, the LV was the second best, and PPI was the worst. In some sample sets, predictions with the PPI distance produced nearly random accuracies (0.5).

Table 3 Cross validation accuracies with various combinations of protein features[a]

Data composition	CV accuracy (AUC)	FP rate
(SMI, SID) & (PPI, SEQ, LV)	81.35% (+5.36% than PPI, SEQ)	0.240
(SMI, SID) & (PPI, SEQ)	76.0%	0.258
(SMI, SID) & (SEQ, LV)	79.66% (+3.66% than PPI, SEQ)	0.260
(SMI, SID) & (PPI)	63.44% (3rd)	0.393
(SMI, SID) & (SEQ)	75.87% (1st)	0.287
(SMI, SID) & (LV)	69.88% (2nd)	0.330

[a] Abbreviations in the first column represent included features. (SMI: SMILES, SID: side effects, SEQ: amino acid sequence, PPI: distance in PPI network, LV: localization vector)

2.7. Effects on novel prediction using clinical trial data

To evaluate the effects of the LV on finding unknown relationships between drugs and targets, we collected new data independent of the gold-standard drug-target data used in the CV. We downloaded clinical trial data from ClinicalTrials.gov (http://clinicaltrials.gov), and chose 8644 associations that did not overlap with our known dataset obtained from DrugBank. We treated these drug–target associations as unknown associations that should be detected. Because the clinical trial data contained only drug–disease relationships, we combined it with gene–disease relationships obtained from the OMIM database[33] and then converted it to drug–target relationships. If a disease contained multiple associated proteins, multiple drug–target relationships were developed. From that point, we selected the drug–target associations containing all five features used in the CV. After preprocessing, 103 associations remained, which included 83 drugs and 16 diseases.

We applied the ten classifiers, which were trained as described in Section 2.7 for these novel associations. Each classifier predicted the classes for the associations and we applied majority voting with the CV accuracy as their weight. As a result in Table 4, the classifiers with (PPI, Seq, LV) predicted 43 associations (28 + 15); however, the classifiers with (PPI, Seq) predicted 22 associations (15 + 7). Notably, the classifier with the LV could predict 28 unknown associations (Table 5), which were missed when the LV was excluded. Even though 7 associations were missed with the LV, the obtained advantages were greater than the disadvantages.

When we compared prediction results from the (LV, Seq) and (PPI, Seq) datasets, the prediction result with (LV, Seq) contained 1.5 times as many associations as in the (PPI, Seq) prediction result.

Table 4 Summary of the novel prediction results with and without LVs[a]

Experiment	#With-LV	#Common	#Without-LV
(PPI, Seq, LV) *vs.* (PPI, Seq)	28 (PPI, Seq, LV)	15	7 (PPI, Seq)
(LV, Seq) *vs.* (PPI, Seq)	23 (LV, Seq)	10	12 (PPI, Seq)

[a] The first column describes configurations for each experiment. The second column represents the number of associations found only in the prediction including LVs. The third column shows the number of associations found in both predictions, and the last column shows the number of associations found only in prediction that excluded LVs. The drug features, SMILES and side effects were included in all prediction experiments.

Table 5 List of unknown associations found only in the experiment with (PPI, Seq, LV)

Predicted disease	Drugs (clinical trial identifiers)
Hypertension	*Aliskiren* (NCT01184599), *amlodipine* (NCT00558064, NCT00558428, NCT00860262), *carvedilol* (NCT02056626), *epoprostenol* (NCT00004754), *hydrochlorothiazide* (NCT00000525), *losartan* (NCT00168857), *metoprolol* (NCT00060918, NCT00060931), *sibutramine* (NCT00679653), *telmisartan* (NCT00168857, CT00550953, NCT00599885, NCT00860262), *trandolapril* (NCT00235014), *valsartan* (NCT00171119, CT00599885), *verapamil* (NCT00235014)
Attention-Deficit/ Hyperactivity Disorder (ADHD)	*Atomoxetine* (NCT00252278), *methylphenidate* (NCT01130467)
Migraine	*Eletriptan* (NCT00259649), *naratriptan* (NCT01726920)
Heart disease	*Ezetimibe* (NCT00639158), *triamterene* (NCT00000525)
Mental retardation	*Bromocriptine* (NCT00004300), *sertraline* (NCT00491478)
Dysplasia	*Cyclophosphamide* (NCT00322101), *cyclosporine* (NCT00322101, NCT01231412), *etoposide* (NCT00602771), *hydrocortisone* (NCT00004669), *methotrexate* (NCT00322101, NCT01789255), *tacrolimus* (NCT00322101), *vorinostat* (NCT01789255)
Obesity	*Salsalate* (NCT00258115)

Table 6 Summary of the results with propagated/raw LVs

Experiment	Propagated LVs	Raw LVs
CV accuracy (AUC)	0.8135	0.8138
FP rate	0.240	0.234
# of found unknown associations	43	38
# of associations found only in this experiment	6 Dysplasia: *cyclophosphamide*, *cyclosporine*, *dexamethasone*, *hydrocortisone*, heart disease: *Ezetimibe* malaria: *quinine*	1 hypertension: *Timolol*

2.8. Effects of LV propagation

Finally, the effect of propagation is evaluated in this section. As many drug targets and candidate proteins have their LVs, a candidate association can be examined with more diverse evidence. This is a purpose of the propagation, and one of the main contributions of this study.

Before the propagation, 15.63% of proteins and 94.41% of known drug targets contained nonzero LVs. After the propagation had been completed, 18.45% of proteins and 98.89% of known drug targets had nonzero LVs. Even when we merged two popular PPI databases, the PPI network was separated; therefore, the propagation was restricted. Nevertheless, the propagation showed a promising result in finding unknown associations. To test the effectiveness of the propagation, we run again the experiments described in Sections 2.6 and 2.7 with propagated LVs and raw LVs. The results are summarized in Table 6. The CV accuracy and FP rate were similar in two experiments. In the prediction test, however, 6 associations were found only with propagated LVs. Even though only 5% of proteins get LVs through the propagation, 10% more associations were found. This result supports our expectation that the propagation help increasing the coverage of the localization information. If the amount and quality of PPI data are advanced sufficiently, we could expect better propagation effects.

3. Conclusion

In this study, we proposed a novel method for exploiting protein localization information as a feature for computational drug repositioning. We organized the localization information into a localization vector LV, and remedied the incompleteness of the LV by network propagation. The LV has an informative pattern for predicting a given drug target; it shows independent characteristics compared to existing features. We determined that when LVs were included, a prediction algorithm produced better results in a cross-validation experiment. While predicting unknown associations from real clinical trial data, the inclusion of LVs improved the prediction coverage. The propagation of LVs increased the coverage of the prediction result compared to a result using raw LVs.

4. Methods

4.1. Localization vector propagation

Not all proteins have their known localizations; therefore, we inferred unknown protein localizations to increase the LV coverage. Based on an assumption that physically interacting proteins have similar subcellular localizations, we iteratively propagated LVs through the PPI network. A protein takes the LVs of its neighbor and merges the incoming LVs with its own initial LV in each iteration. A LV v at time t for a protein is defined as:

$$v_t = \alpha v_{\text{incoming}} + (1 - \alpha)v_0$$

where v_0 represents the initial LV of the protein, α is a diffusion factor, and v_{incoming} is the sum of the LVs of the neighbor proteins. This propagation was iteratively run until the following convergence condition is satisfied.

$$\left| \sum_{u \in \text{LV}_t} \text{norm}_{\max}(u) - \sum_{v \in \text{LV}_{t+1}} \text{norm}_{\max}(v) \right| < 10^6$$

where the $\text{norm}_{\max}(u)$ of a LV u is defined as $\max(|u_1|, |u_2|, \ldots, |u_{10}|)$ when $u_1, u_2, \ldots, u_{10}$ represent attributes of u.

Let us define LV_p as a set of the propagated LVs and LV_0 as a initial set of LVs. In the LV_p, more proteins had their localizations; however, known localization information was modified. Finally, we merged the LV_0 and the LV_p to generate the final LVs based on an assumption that known localization information is more reliable than the propagated localization information. Thus, we copied the LV of a protein from the LV_p to the LV_0 if the protein had no localization information in the LV_0.

The diffusion factor a affects the propagated distance of the LV information through the PPI network. If a high, the LV information reaches farther proteins. If a is low, a protein without the LV will receive the information from the nearest proteins. Accordingly, we evaluated various values for a from 0.1 to 0.9 using the method described in Section 2.5. As a result, we found that a has only a minor effect on the distribution of the LV. In this study, we set a as 0.4.

4.2. Prediction method for cross-validation and novel rediction

The LV is a kind of prediction feature that can be applied for many prediction algorithms. In this paper, we employed and modified a pipeline of PREDICT,[15] which is one of the start-of-art prediction algorithms exploiting heterogeneous features. PREDICT is a flexible approach for applying many combinations of features. To predict a drug–target association as true or false, a scoring scheme for the association must be defined. When there are known drug–target associations (d,t) and a query drug–target association (d′,t′), a score for the query association S(d′,t′) is calculated according to the most similar (d,t), as follows:

$$S(\mathrm{d}', \mathrm{t}') = \max \sqrt{(\mathrm{sim}(\mathrm{d}, \mathrm{d}') \times \mathrm{sim}(\mathrm{t}, \mathrm{t}'))}$$

The function sim( ) is defined as the average of similarities of non-zero features for the pair. For example, when two proteins t_1 and t_2 have sequence similarity 0, PPI distance 0.4, and LV similarity 0.6, the sim(t_1, t_2) = (0.5 + 0.5)/2 = 0.5.

In this prediction pipeline, known drug–target data is necessary. Therefore, we divided the known drug–target associations into ten subsets and performed ten runs for build classifiers. In each run, one of the ten subsets was considered as the query drug–target association (positive samples); the same number of random drug–target pairs was generated as negative samples. The remaining nine subsets were considered as known drug–target associations for calculating S(d′,t′) for the positive and negative samples. After calculating the score for the samples, we applied a ten-fold CV to train a logistic regression classifier and investigated whether the classifier could distinguish two classes by the given scores. If features in calculating the scores were effective, a more clear separation between the two samples would be produced.

The ten trained classifiers were exploited in the prediction. When we predicted unknown associations from clinical trials, we used the entire known drug–target data for calculating scores for the unknown associations. Then, ten classifiers performed their predictions, and a final prediction was made by majority voting. Accuracies of the models in the CV were exploited as weights for the models in the majority voting.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2015R1A2A1A05001845)

Notes and references

1 F. Pammolli, L. Magazzini and M. Riccaboni, *Nat. Rev. Drug Discovery*, 2011, **10**(6), 428–438.
2 Z. Liu, H. Fang, K. Reagan, X. Xu, D. L. Mendrick, W. Slikker and W. Tong, *Drug Discovery Today*, 2013, **18**(3), 110–115.
3 S. Ekins, A. J. Williams, M. D. Krasowski and J. S. Freundlich, *Drug Discovery Today*, 2011, **16**(7), 298–310.
4 D. Sardana, C. Zhu, M. Zhang, R. C. Gudivada, L. Yang and A. G. Jegga, *Briefings Bioinf.*, 2011, **12**(4), 346–356.
5 F. Iorio, T. Rittman, H. Ge, M. Menden and J. Saez-Rodriguez, *Drug Discovery Today*, 2013, **18**(7), 350–357.
6 W. Loging, R. Rodriguez-Esteban, J. Hill, T. Freeman and J. Miglietta, *Drug Discovery Today: Ther. Strategies*, 2012, **8**(3), 109–116.
7 J. Lamb, *Nat. Rev. Cancer*, 2007, **7**(1), 54–60.
8 C. Wu, R. C. Gudivada, B. J. Aronow and A. G. Jegga, *BMC Syst. Biol.*, 2013, **7**(suppl. 5), S6.
9 F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu and Y. Tang, *PLoS Comput. Biol.*, 2012, **8**(5), e1002503.
10 S. Zhao and S. Li, *Bioinformatics*, 2012, **28**(7), 955–961.
11 G. Jin, C. Fu, H. Zhao, K. Cui, J. Chang and S. T. C. Wong, *Cancer Res.*, 2012, **72**(1), 33–44.
12 H. Zhao, *et al.*, *Cancer Res.*, 2013, **73**(20), 6149–6163.
13 P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon and V. Mooser, *Nat. Biotechnol.*, 2012, **30**(4), 317–320.
14 Z. Y. Wang and H. Y. Zhang, *Nat. Biotechnol.*, 2013, **31**(12), 1080–1082.
15 A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan, *Mol. Syst. Biol.*, 2011, **7**, 496.
16 C. E. Au, A. W. Bell, A. Gilchrist, J. Hiding, T. Nilsson and J. J. M. Bergeron, *Curr. Opin. Cell Biol.*, 2007, **19**(4), 376–385.
17 A. Zhang, H. Sun and X. Wang, *BMC Syst. Biol.*, 2012, **6**, 20.
18 P. Cahwley, H. B. Samal, J. Prava, M. Suar and R. K. Mahapatra, *Genomics*, 2014, **103**, 83–93.
19 S. Telkar, H. S. S. K. Kumar and R. Mahmood, *Star Journal*, 2013, **2**(4), 34–39.
20 S. Park, J. S. Yang, Y. E. Shin, J. Park, S. K. Jang and S. Kim, *Mol. Syst. Biol.*, 2011, **7**, 494.
21 M. Deng, K. Zhang, S. Mehta, T. Chen and F. Sun, *J. Comput. Biol.*, 2003, **10**(6), 947–960.
22 D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Res.*, 2008, **36**(suppl. 1), D901–D906.

23 M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, *Mol. Syst. Biol.*, 2010, **6**, 343.
24 M. Magrane and UniProt Consortium, *Database*, 2011, bar009.
25 K. R. Brown and I. Jurisica, *Genome Biol.*, 2007, **8**(5), R95.
26 A. Chatr-aryamontri, *et al.*, *Nucleic Acids Res.*, 2013, **41**(D1), D816–D823.
27 R. C. Team, *R Foundation for Statistical Computing*, 2012.
28 G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang, *Bioinformatics*, 2010, **26**(7), 976–978.
29 P. Resnik, *J. Artif. Intell. Res.*, 1999, **11**, 95–130.
30 T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 1981, **147**(1), 195–197.
31 L. Perlman, A. Gottlieb, N. Atias, E. Ruppin and R. Sharan, *J. Comput. Biol.*, 2011, **18**(2), 133–145.
32 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *ACM SIGKDD explorations newsletter*, 2009, **11**(1), 10–18.
33 A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic Acids Res.*, 2005, **33**(suppl. 1), D514–D517.

