Network-based drug-disease relation prioritization using ProphNet

Víctor Martínez, Carmen Navarro, Carlos Cano, and Armando Blanco

Department of Computer Science and AI, University of Granada.
fvictor@correo.ugr.es,cnluzon@decsai.ugr.es,
ccano@decsai.ugr.es,armando@decsai.ugr.es

Abstract. Assisting drug repositioning processes can lead to a considerable reduction in cost and time in any drug development process. Recent *in silico* approaches have addressed the network-based nature of biological information to assess the possible new indications for a query drug. Here we present a new methodology based on network prioritization, that can aid researchers in the drug repositioning process by means of prioritizing drugs related to a query disease. Results show that selection of the data sources to be integrated can be a critical step towards success in drug repositioning.

Keywords: drug repositioning, prioritization, disease networks, data integration

1 Introduction

Developing a new drug is a risky process, now estimated to last about 15 years and cost between $800 million and $1 billion [1]. This amount can be considerably reduced if an already commercialized drug is used for new indications. This task is known as drug repositioning, and there are several classic examples of large benefits produced by a successful drug repositioning, such as Viagra or Minoxidil [2].

In order to reduce time and resources needed for drug development, some efforts have been made for *in silico* drug repositioning (for a review, see [1]). These could be encompassed in two main categories: those focused on composition, chemical or molecular features of drugs, and those based on knowledge about diseases, their underlying processes or their symptomatology. Regarding the first category, methods may relate drugs based on quantitative chemical measures from both drugs and targets [3, 4].

On the other hand, disease-focused proposals try to relate drugs and diseases based on symptomatology, known treatments or pathological information. This category includes approaches such as Chiang and Butte's application of the 'Guilt-by-association' principle [5], which assumes that two diseases are related when a similar treatment is prescribed for both (i.e. prescriptions share a considerable subset of drugs). Approaches like Promiscuous [6] take into account side-effects information to relate drugs and diseases. Both diseases and

side-effects imply certain symptomatology or underlying biological activity, and their phenotypic expressions are usually similar. However, same phenotypic expression might have different possible underlying causes.

Nonetheless, relations between biological entities are complex and mostly network-structured, and it is not yet clear which among these approaches is best, if any [7]. Symptomatology-based approaches seem to suit better when there is a lack of knowledge about the molecular processes underlying the query disease. On the other hand, molecular-based proposals should be used when there is expertise in certain target's chemical behaviour. Integrating several data sources seems a solution. However, certain sources such as expression data or pathway information are yet scarce or difficult to obtain. Data integration should be done carefully, since redundant, contradictory or vague information could lead to poor results. Consequently, best data source configuration appears to be disease-dependent. Therefore, deciding which sources of data should be analysed in advance is a difficult task.

In this work, we propose a new methodology for drug repositioning based on networks prioritization. Different possible configurations of data sources are studied in order to test its performance prioritizing diseases against sets of drugs. Results show that integrating only adequate knowledge for drug repositioning can draw promising results while using other types of data such as pathways, although promising, is not efective since they are not yet sufficiently mature to be used in the process of prioritization.

2 Methodology

We have applied ProphNet [8] to prioritize drugs and diseases. ProphNet is a general network-based prioritization tool which has shown excellent results for gene-disease prioritization in previous works [8].To apply ProphNet to drug repositioning, we first need to define and build the data networks the algorithm is applied to. This representation considers one network for each type of entity (e.g. one network modelling gene-gene interactions, one for drug interactions, etc.). Each network node v represents a biological entity (e.g., drug or disease) labelled with a value $\Psi(v)$. Nodes in networks are connected by weighted arcs representing an interaction or relationship between the connected pair of nodes. There are two types of networks: networks which represent relations or interactions between elements from the same domain and networks which represent relations or interactions between elements from two different domains. Network connections are represented as adjacency matrices. Each adjacency matrix A is normalized as

$$A_{norm} = D^1_G * A * D^2_G,$$

where D^1_G and D^2_G are diagonal matrices where each component is defined as

$$D^1_{G_{jj}} = 1/\sqrt{(\sum_{k=1}^{c} A_{jk})}\ j = 1, .., r$$

$$D^2_{G_{kk}} = 1/\sqrt{(\sum_{j=1}^{r} A_{jk})}\ k = 1, .., c.$$

These networks are used to build a Global Graph which is a network containing all other previously defined networks. Our goal is to measure the degree of relation between two sets of nodes (called Query Set and Target Set, respectively) from two different networks (called Query Network and Target Network, respectively). The Query Set Q is provided by the user as input (e.g. a set of drugs or diseases of interest) while the Target Set T is iteratively established by ProphNet (to find out the most strongly related diseases or drugs, respectively). Nodes in Q are initially set to: $\Psi(v) = 1/|Q|\ \forall v \in Q$ and all the nodes v from T are initially set to: $\Psi(v) = 1$. The rest of the nodes are set to zero.

We define a path connecting the Query Network and the Target Network as a path of networks (not a path of nodes) which allows to get from Q to T. Two propagation operations are defined: "propagation within a network" and "network-to-network propagation". First operation allows to propagate node values within a specified network using the Propagation Flow algorithm [9, 10]. This algorithm is performed by iteratively applying

$$x_{i+1} = (1 - \alpha) * M * x_i + \alpha * x_0,$$

where α is a parameter which determines the importance of the prior information in the network, M is the normalized adjacency matrix of the network and x_i is a vector representing network node values at iteration i. The second operation allows to propagate values from the current network to the following network in the path by assigning to each node v from the following network a value computed as

$$\Psi(v) = \frac{\sum^{x \in neig(v)} \Psi(x)}{|neig(v)|},$$

where $neig(v)$ is the set of nodes from the current network which are connected with node v in the following network.

Initially, node values in the Query Set are propagated within the Query Network using the "propagation within network" operation. The same process is performed in the Target Network to propagate values from the Target Set nodes. Then, for all the possible paths from the Query Network to the Target network, values are propagated using the two mentioned operations alternatively until all the networks adjacent to the Target Network are reached. Finally, vectors representing adjacent networks' node values for each path are multiplied by the normalized adjacency matrix of the network connecting the adjacent network with the target network and the resulting vectors are correlated simultaneously with a vector representing the Target Network node values (using Pearson Correlation). To perform this correlation simultaneously, resulting vectors are concatenated in only one vector and are correlated with a vector obtained by concatenating a vector representing Target Network node values as many times as the number of paths. The computed correlation value is used as a score to determine the degree of relationship between the Query and Target sets.

In order to score each entity in the Target Network according to the degree of relationship to the Query Set, each node from the Target Network is iteratively set as Target Set and its score is computed using the method described above.

Finally, prioritized lists are obtained by sorting all the target node scores in decreasing order.

Since our method can be used with different network configurations, we have tested ProphNet with different global graph options to select the best configuration for drug repositioning (see Results). Three options have been considered (see Figure 1). Finally, a validation test using the best configuration has been performed to prioritize clinical trials obtained from ClinicalTrials.gov as described in Materials.

3 Materials

The disease phenotype network has been derived from the Online Mendelian Inheritance in Man database (OMIM,[11]) using text-mining as described by [12]. A profile has been created for each phenotype by counting the number of appearances of some MeSH vocabulary terms. Cosine distance has been computed for each phenotype pair to build an interconnected network of diseases. Finally, only the strongest relations are maintained. 5080 disease phenotypes with 39458 weighted relations were extracted.

The drug network has been extracted from DrugBank [13]. Two drugs are connected by an arc in the network if at least one interaction between them is found in DrugBank. Only drugs with at least one interaction are considered, obtaining 1109 unique drugs connected by 10906 interactions.

The protein domain network has been derived from DOMINE [14] and InterDom [15] revealing 48778 unique relations between 5490 protein domains. 1614 protein domain-drug relations were extracted from Pfam [16] and from annotations of nsSNPs in the UniProt database [17].

The protein (gene) network was obtained from Human Protein Reference Database (HPRD,[18]). This protein-protein interaction network contains 64662 unique interactions between 8919 proteins. The drug-protein interactions were also extracted from DrugBank (2860 interactions). Gene-disease relations were directly extracted from OMIM (1393 relations). Finally, gene-protein domain relationships were extracted from Pfam entries.

The drug-disease network has been computed by mapping disease names to UMLS concepts and matching these with drugs indications from DailyMed as described by [3]. 1337 drug-disease relations were obtained.

Clinical trials data were extracted from ClinicalTrials.gov. Only clinical trials with drugs and diseases in our datasets were considered, obtaining 1632 drug-disease relations under study. The phase for each clinical trial was also obtained.

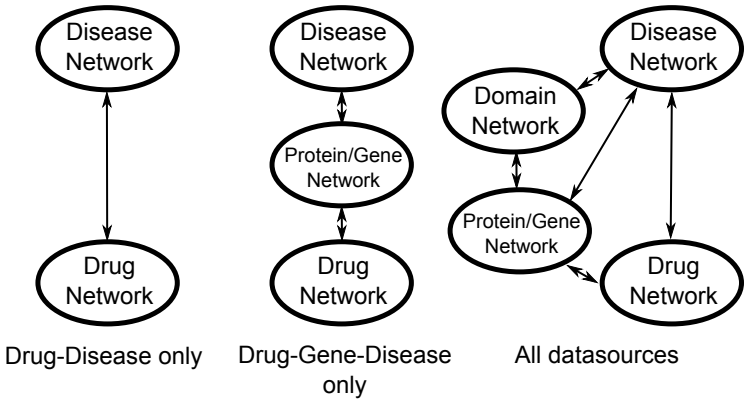

Fig. 1. Tested networks configurations.

4 Results and Discussion

Validation tests were applied in order to select the best network configuration and measure the performance of our approach. First, a leave-one-out (LOO) test was performed for each global graph configuration obtaining its performance. After selecting the best configuration, a test using this configuration was performed to prioritize relations obtained from clinical trials.

4.1 Best network configuration tests

Leave-one-out validations were performed to determine the best data sources for the drug repositioning task. LOO tests consists of 1337 test cases (one for each explicit drug-disease relationship in the global graph). A leave-one-out prioritization was performed for each test case by removing one known drug-disease relation, taking the drug as query set and checking the resultant disease ranking to measure performance. Receiver operating characteristic (ROC) curves have been plotted for each LOO validation test. A ROC curve is created by plotting the fraction of true positives out of the positives vs. the fraction of false positives out of the negatives at various threshold settings. A true positive occurs when the rank of the case disease is below the threshold. A false positive occurs when a disease that is not in the case is ranked below the threshold. The area under the ROC curve (AUC value) was also computed to quantify gains. Finally, the mean ranking position of the query disease in the prioritized lists obtained for each test case was also computed and normalized by dividing by the total number of elements in the list (5080 diseases).

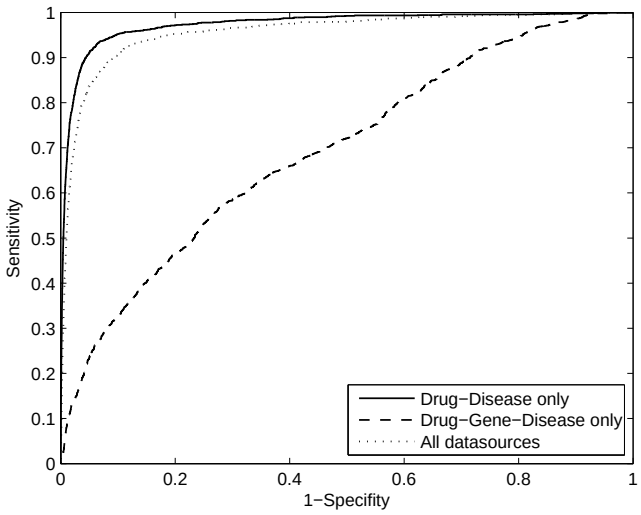

Fig. 2. ROC curves for different data source combinations.

We measured the accuracy of the ranking for three different data configurations (Figure 2). The tests with all the data sources obtained a 0.9564 AUC value and 223 ± 566 mean ranking. Tests with only drug-protein-disease networks obtained a 0.695 AUC value and 1550 ± 1370 mean ranking. Tests with only disease-drug networks obtained a 0.9738 AUC value and 134 ± 438 mean ranking. Therefore, our method achieves the best performance in drug repositioning when only disease-drugs relationships are considered.

4.2 Clinical trials validation test

To validate the results obtained by the best network configuration for drug repositioning in real cases, we applied Prophnet on this "Drug-Disease only" network to prioritize relations derived of clinical trials from ClinicalTrials.gov as described in Materials. Drug-disease relations from these clinical trials were removed from our global network if already explicitly present in our data in order to perform a blind prioritization. A 0.9288 AUC value was obtained for the whole dataset, proving the high performance of this approach in real cases.

Table 1 summarizes the results obtained. Drugs in earlier stages of clinical trials have a high risk of failure due to toxicity or lack of efficacy. As can be seen in the table, results are better for clinical trials in more advanced phases. Therefore, suggested drugs repositioned by our approach are more likely to succeed. Although the coverage in each stage may seem low, our approach successfully predicts almost one out of four clinical studies in their last stage of development. Predictions made are therefore reliable and have the potential to reduce costs

considerably, thus we consider they may be of interest to the pharmacological industry.

Clinical trials phase	Case count	AUC	% in Top 20
N/A	267	0.9067	18.73
Phase 0	9	0.9532	11.11
Phase 1	144	0.9347	9.03
Phase 2	495	0.9429	9.29
Phase 3	377	0.9165	18.30
Phase 4	340	0.9363	24.12
All phases	1632	0.9288	15.99

Table 1. Results obtained for drug-disease prioritization of clinical trials recently performed or currently under development. First column shows the phase of the study, second column the number of studies in this phase, third column AUC value obtained and fourth column the percentage of cases ranked in top 20.

5 Conclusions

A new methodology for *in silico* drug repositioning has been presented in this work. We have based our approach in two main ideas. Firstly, biological entities interact with each other in a networked, intricate way. Consequently, any element should be observed as a connected entity interacting with its environment, rather than as an isolated element. Furthermore, biological information is diverse and growing. Our data source study has shown that data sources selection and configuration can be a critical step, since redundant or vague information can deteriorate otherwise good results. Eventually, the simplest and most reduced data source selection, taking into account only diseases and drugs, and the known interactions between them, was the one drawing better results, and therefore the one best suited for drug repositioning.

Results have shown that this approach can elucidate unknown drug applications with a high level of confidence in real situations, therefore these methods may potentially save a large amount of resources in the drug development process.

Acknowledgments. This project is part of projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid.

References

1. Dudley, J., Deshpande, T., Butte, A.: Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform **12**(4) (2011) 303–311
2. Ashburn, T., Thor, K.: Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov **3**(8) (2004) 673–683

3. Gottlieb, A., Stein, G., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol **7**(1) (2011)
4. Li, Q., Cheng, T., Wang, Y., Bryant, S.: Pubchem as a public resource for drug discovery. Drug Discov Today **15**(23) (2010) 1052–1057
5. Chiang, A., Butte, A.: Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Int J Clin Pharmacol Ther **86**(5) (2009) 507–510
6. von Eichborn, J., Murgueitio, M., Dunkel, M., Koerner, S., Bourne, P., Preissner, R.: Promiscuous: a database for network-based drug-repositioning. Nucleic Acids Res **39**(suppl 1) (2011) D1060–D1066
7. Swinney, D., Anthony, J.: How were new medicines discovered? Nat Rev Drug Discov **10**(7) (2011) 507–519
8. Martínez, V., Cano, C., Blanco, A.: Network-based gene-disease prioritization using prophnet. In: NETTAB 2012 (2012)
9. Vanunu, O., Sharan, R.: A propagation based algorithm for inferring gene-disease associations. In: Proceedings of German Conference on Bioinformatics, Citeseer (2008)
10. Navlakha, S., Kingsford, C.: The power of protein interaction networks for associating genes with diseases. Bioinformatics **26**(8) (2010) 1057–1063
11. Amberger, J., Bocchini, C., Scott, A., Hamosh, A.: Mckusick's online mendelian inheritance in man (omim®). Nucleic Acids Res **37**(suppl 1) (2009) D793–D796
12. Van Driel, M., Bruggeman, J., Vriend, G., Brunner, H., Leunissen, J.: A text-mining analysis of the human phenome. Eur J Hum Genet **14**(5) (2006) 535–542
13. 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 **39**(suppl 1) (2011) D1035–D1041
14. Raghavachari, B., Tasneem, A., Przytycka, T., Jothi, R.: Domine: a database of protein domain interactions. Nucleic Acids Res **36**(suppl 1) (2008) D656–D661
15. Ng, S., Zhang, Z., Tan, S., Lin, K.: Interdom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res **31**(1) (2003) 251–254
16. Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., et al.: The pfam protein families database. Nucleic Acids Res **38**(suppl 1) (2010) D211–D222
17. Magrane, M., et al.: Uniprot knowledgebase: a hub of integrated protein data. Database (Oxford) **2011** (2011)
18. Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, et al.: Human protein reference database—2009 update. Nucleic Acids Res **37**(suppl 1) (2009) D767–D772