Network-based approach for drug repurposing using drug signature and disease phenotype

Ariane Alix
Department of Mathematics
École Normale Supérieure Paris-Saclay
94230 Cachan, France
ariane.alix@ens-paris-saclay.fr
Supervisors: Clémence Réda and Emilie Kauffman

Abstract

Drug repurposing consists of the investigation of already used drugs, to see if they can be used for treating other diseases. This approach has recently raised thanks to the access to large-scale perturbation databases, such as the Library of Integrated Network-based Cellular Signatures (LINCS) and the development of new computational methods. In this paper, we review a drug–disease associations prediction method based on a recommendation system (bipartite graph) where the studied features are the drug signatures and diseases phenotypes.

1 Introduction

We consider two types of gene expression profiles: disease phenotype or signature (can be used interchangeably) and drug signature. A disease phenotype is usually a list of genes, that are up or down-regulated when compared to healthy controls. A drug signature is the gene expression response, within cells, when they are stimulated with a drug.

Approaches based on this type of data can lead to drug-repurposing techniques. For example, a method described in [4] directly compares diseases and drugs signatures to infer repurposing opportunities. Concretely, the comparison of drug and disease gene expression profiles can be used to generate hypotheses of drug-repurposing when we observe a negative correlation between the profile of a drug and of a disease: the drug "counter-acts" the effects of the disease on the genes.

In our approach, we will use the drug and diseases signatures to calculate drug-drug and disease-disease similarity. Then, we will compute a recommendation system based on known drug-disease associations and those similarities to compute drug recommendations for the diseases in our database. Our work will be based on Alaimo's drug-target recommendation system [13], the difference being in the utilization of full diseases data, and the computation of similarity metrics using gene signatures. We will first review some similarity metrics and associated methods for drug and diseases studies, then describe our implementation and discuss our results.

2 Review of some methods used for drug-disease associations

2.1 Computation of drugs signatures and diseases phenotypes

We will review two statistical methods used to computationally extract drug signatures and disease phenotypes from messenger RNA expression data: the MODZ method and the CD method which is more recent. Concretely, these "signatures" and "phenotypes" are quantified profiles representing the impact of drugs or diseases of the genes. Table 1 is an example of what drugs signatures can look like:

Table 1: Example of drugs signatures

	Genes		
Drug	RNF14	UBE2Q1	RNF17
Methyl 2,5-dihydroxycinnamate	0.009656854	0.007538027	0.003181712
Compound 10	0.011227879	-0.003624232	-0.002378904
GSK-3-inhibitor-II	0.013539515	-0.003986923	-0.013635419

2.1.1 Moderated Z-score method

The moderated Z-score (MODZ) is a statistical method that was used to compute gene expression profiles; for example the impact of diseases on the genes in human cell lines. To generate these profiles, the MODZ method looks at experiments done in some biological replicates (parallel measurements of biologically distinct samples that capture random biological variation), and computes a score as a weighted average of the replicates signatures where the weights are proportionnal to their Spearman correlations.

This method is the one that was used to compute LINCS L1000, a comprehensive database of more than a 1.3 million gene expression profiles [1]. We will use data of drug signatures extracted from this database for the implementation of our algorithm later on.

2.1.2 Characteristic Direction method

The Characteristic Direction (CD) is a geometrical approach used to describe the impact of drugs and diseases on gene expression. The method has been studied by Clark in 2014 [3]. The main differences with the MODZ method is that it gives less importance to large changes of individual genes, and more importance to large groups of genes moving in the same direction even if the change is of smaller amplitude. Its has been proven to be more accurate than previous methods for the identification of differentially expressed genes ([2]).

This new approach to quantify gene expression profiles had a significative impact on the computation of drugs and diseases signatures. In particular, it has been used in 2016 by Qiaonan Duan, St Patric Reid et al. on the LINCS L1000 database to create a search engine called L1000CSD^2 [2]. This search engine provides a database of thousands of small-molecules and their signatures, and allows to find which gene expression profile it can mimic or reverse. As a consequence, it can predict drug targets by looking for the molecules that would reverse the effect of a disease.

2.2 Connectivity Mapping

The Connectivity Mapping is an approach on which are based numerous methods for drug discoveries studies. In particular, it can be used to discover correlations between disease and drug gene expression, and thus help find drug-disease associations [4].

2.2.1 Standard CMap methodology

The basic concept of CMap is to use a reference database containing drug-specific gene expression profiles and compare it with a disease-specific gene signature [8].

Concretely, a gene signature is made of scores representing the effect on the genes of the studied drug or disease, and those scores are ranked from the most 'upregulated' gene to the most 'downregulated'. From there, we compute a 'connectivity score' (the similarity metric) between drugs or diseases, classically with a Kolmogorov-Smirnov statistic on the two lists of upregulated and downregulated genes scores. The process for a drug profile is summarized in Figure 1.

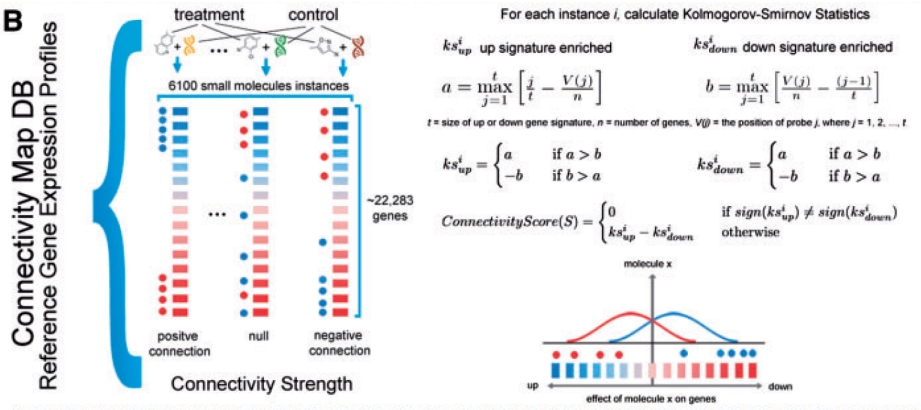

Figure 1: Connectivity Score computation

With this principle, a negative connectivity score between the profile of a disease and a drug signature means that the exposure to that drug could reverse the expression pattern of the disease (hence are likely to "cure" it).

2.2.2 Variation of similarity metric: ssCMap

The Statistically Significant Connectivity Map (ssCMap) is a variation of the standard CMap methodology. It has been developed in 2008 by Zhang et al. [5], and introduces a new ranking score that aims to solve an issue with the KS similarity metric of the standard version: a top-ranked drug could have strong effects on a subset of functions associated to the disease, but also affect other functions (hence important side-effects).

Let's assume that we want to compute the connectivity score between a drug with a signature of m genes, and a disease with a phenotype of N genes ($N \geq m$). For the drug, the ith most important gene effect will be assigned the rank $(m - i + 1)$ if it is upregulated and $-(m - i + 1)$ if it is donwregulated. The same principle applied to the disease. It can also be done with two drugs or two diseases. We can compute the connectivity score between their two gene expression profiles with the following formula:

$$C(R_1, R_2) = \frac{\sum_{i=1}^{m} R_1(g_i) R_2(g_i)}{\sum_{i=1}^{m} (m - i + 1)(N - i + 1)}$$

Where g_i is the ith gene of the signatures, and $R_1(g_i)$, $R_2(g_i)$ are its respective signed ranks in the signatures of the drug and the disease. The numerator is the 'connection strength', the denominator is the 'maximum connection strength'.

The score ranges from -1 to 1, which is easily interpretable: 1 means that the two profiles have the same biological effect, -1 that they have the opposite effect (hence a cure in the case of a drug-disease association).

Since this method is easily implementable (more than other methods like CMapBatch considering multiple gene phenotypes of a same disease), and robust, we will test this similarity metric in our network implementation.

2.2.3 Variation of similarity metric: XCos

The eXtreme cosine (XCos) method is an alternative similarity metric using the Anatomical Therapeutic Chemical (ATC) classification. The principle of the method is to keep only the top N and bottom N genes affected in a signature (most upregulated and downregulated), and to set the score of the others to 0. The similarity (connectivity score) is then computed with a dot product of the signatures. It is therefore similar to the ssCMap version, except that the middle-range gene scores are set to 0.

It was studied by Cheng et al. [6] in 2013, who showed that it could outperform the standard CMap when it was with a large number of genes (number of top genes selected to compute the connectivity

scores). It is also good at determining drug compound classes, which could be useful in a graph implementation such as the one we will do, and is robust to false positives which is good in the context of drug repurposing.

3 Implementation

3.1 Recommendation systems

A recommendation system is a system of users and objects, where each user collects some objects, for which he can also express a degree of preference [13]. We can represent this system as a bipartite graph. The purpose of a recommendation algorithm is to find which new objects could appeal to an user, and give scores proportional to the likeliness of the match.

In the case of Alaimo's implementation, the users are the drugs and the objects are the proteins targeted. In our implementation, based on Alaimo's, the users would also be the drugs but the objects would be the diseases.

3.2 Alaimo's drug-target recommendation network

We remind that a "drug target is a molecule in the body, usually a protein, that is intrinsically associated with a particular disease process and that could be addressed by a drug to produce a desired therapeutic effect" (cf. `https://www.sciencedirect.com/topics/chemistry/drug-target`). Alaimo's method aims to predict possible new drug-target interactions using a network approach.

Our goal later-on will be to adapt this method to the prediction of drug-disease associations using signatures like CMap methods [4], [8].

3.2.1 Object-projection of a bipartite graph

Given a bipartite graph $G(U, O, E)$ (U is the set of users, O of objects and E of edges). We can compute a new graph called "object-projection" such that:

- The nodes are the objects from G,
- There is an edge between two nodes if there exists a path in G through users between these two objects,
- The weights w_{ij} between objects o_i and o_j are computed as a function of the adjacency matrix between users and objects, the orders of the nodes and the similarity matrix between objects and users (computation done similarly to a two steps ressource allocation process).

3.2.2 Calculation of the weights of the object-projection graph

We will consider the context of drug-target prediction ($G(D, T, E)$). Let N_t be the number of targets, N_d the number of drugs. We note:

- $A = \{a_{ij}\}_{N_t \times N_d}$ the adjacency matrix between targets and drugs in graph G,
- $S^t = \{s^t_{ij}\}_{N_t \times N_t}$ the targets similarity matrix,
- $S^d = \{s^d_{ij}\}_{N_d \times N_d}$ the drugs similarity matrix.

We can compute a final similarity matrix between targets, based on the natural similarity of targets and the common interactions via drugs in the network:

$$\begin{aligned} s_{ij} &= \alpha s^t_{ij} + (1-\alpha)\frac{\sum_{k=1}^{N_d}\sum_{l=1}^{N_d}(a_{il}a_{jk}s^d_{lk})}{\sum_{k=1}^{N_d}\sum_{l=1}^{N_d}(a_{il}a_{jk})} \\ &= \alpha s^t_{ij} + (1-\alpha)s^{td}_{ij} \end{aligned}$$

where s^{td}_{ij} describes the similarity between targets t_i and t_j by considering if they are linked by similar drugs, and α is a tuning parameter.

From there we can compute the weights w_{ij} between targets t_i and t_j :

$$w_{ij} = \frac{s_{ij}}{k(t_i)^{1-\lambda} k(t_j)^{\lambda}} \sum_{l=1}^{N_d} \frac{a_{il} a_{jl}}{k(d_l)}$$

where $k(.)$ is the function giving the order of a node in graph G.

In our implementation, we will simply replace the targets with the diseases, and the similarity matrices by our owns.

3.3 Data used

We must note that the different methods of signature computation are compatible regarding the study of their similarity, in the sense that similarity functions look at the ranks of the genes scores, and are therefore not impacted by potential differences in mean or scale of those scores. Indeed, we will use a dataset computed with different methods in the implementation of our algorithm: the MODZ one to compute the drug signatures, and the CD one to compute the disease phenotypes.

The datasets we used are:

- A table of disease phenotypes, with 48 diseases and 11393 genes scores each,
- A table of drug signatures, with 617 drugs and 12717 genes scores each,
- A table of drug-disease associations with 7325 matches of 1543 distincts drugs and 1465 distinct diseases.

All of the 48 diseases of the phenotypes tables are present in the association table. However, only 45 drugs are in both the signature and association tables. Finally, we get 2115 distinct drugs and 1465 distinct diseases when looking at the three tables.

For the association table, we decided to keep only the associations that are marked as 'Approved'. From this table we built the association table A of all the available drugs and diseases (1 if we know that the drug "cures" the disease, 0 otherwise).

The data processing was done in Python, which is very handy for that type of operations.

3.4 Computation

Considering the huge amount of operations for the computations of all matrices (in particular 810 weight matrices for the cross-validation, see 4.1), the computation of all the scores and matrices was done in C++, which is much faster than Python.

3.4.1 Similarity matrices

The first step is to compute the disease similarity matrix and drug similarity matrix based on their signatures similarity. For this part, we used the ssCMap similarity metric (see 2.2.2), and we obtain two matrices: S^{disease} and S^{drug}. After that, we re-scaled the obtained scores between 0 and 1, which has two advantages: the metric becomes distance-like (positivity, symmetricity, triangle inequality), and we can use 0 as a "worst-case scenario" for unknown drugs (which is certainly preferable as a security precaution in the context of drug repurposing). We also added the drugs and diseases for which we do not have a signature or phenotype, by setting their similarity with others to 0 and to 1 with themselves. You can see an example of similarities obtained in Figure 2.

The second step is to compute a similarity matrix between diseases, S^{dd}, based on the interactions with similar drugs in the network (see 3.2.2).

Next, we can compute a final similarity matrix S between disease which is a weighted average depending on a parameter α of the two similarity matrices S^{disease} and S^{dd}.

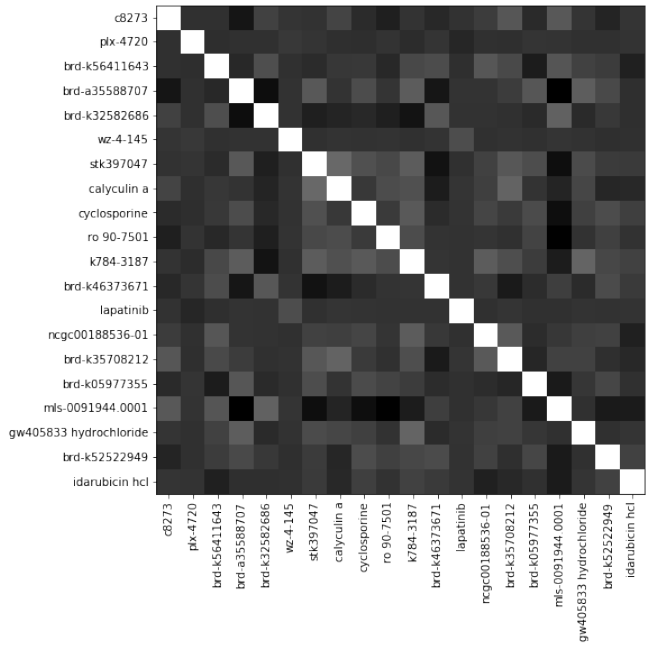

Figure 2: Visualization of the similarity between 20 drugs - Extract from S^{drug}

3.4.2 Weight and recommendation matrix

The final weight matrix is computed exactly as in 3.2.2 using S and depending on a parameter λ. The only step left to obtain the recommendation is to multiply W and A:

$$R = W \times A$$

The obtained R matrix gives the scores of the matches between all drugs and diseases.

4 Results

4.1 Cross-validation

We used cross-validation to assess the ability of the method to recover lost drug-disease associations, and in doing so we chose the best parameters α and λ used to compute S ans W. To that end, we built 10 matrices A^i_{train} that correspond to the matrix A from which we removed randomly 20% of the drug-disease positive associations.

For each of this 10 tables, we computed 9 S^{dd} tables for different values of α; then for each of these S^{dd} tables, we computed 9 W tables for different values of λ. Hence we had to compute 90 S^{dd} tables, 810 W and 81 R tables in total.

We defined two metrics based on the R tables to estimate the "recovery" ability:

- The first metric looks at the TOP X+10 drug scores for each disease, where X is the number of drugs associated to the disease in the full matrix A, and checks if the method ranked the deleted association there:
$$\frac{\text{Number of deleted associations recovered in the TOP X+10 of a disease}}{\text{Number of deleted associations}}$$
- The second metric counts the deleted drug-disease associations that we found again with a score >0:
$$\frac{\text{Number of deleted associations recovered with a score} > 0}{\text{Number of deleted associations}}$$

The best scores are 61.54% and 74.56% for metric 1 and 2 respectively, both obtained with $\alpha = 0.1$ and $\lambda = 0.8$.

4.2 Final table

Using parameters $\alpha = 0.1$ and $\lambda = 0.1$ and the whole table A, we computed a final R matrix giving recommendations of drug-disease associations given all the available data. Four interesting examples:

ind_id	abdominal actinomycosis	previous reco
benzylpenicillin	1.000000	1
phenoxymethylpenicillin	0.845747	0
procaine benzylpenicillin	0.727375	0
cefixime	0.540433	0
cefradine	0.537274	0
cefprozil	0.529992	0
cefaclor	0.529436	0
cyclacillin	0.354836	0
cephalexin	0.328189	0
antipyrine	0.303486	0

ind_id	addison disease	previous reco
fludrocortisone	0.131250	1
hydrocortisone	0.097084	1
cortisone acetate	0.097005	1
prednisone	0.085866	1
dalfampridine	0.081788	0
glatiramer acetate	0.081788	0
betamethasone	0.079831	1
dexamethasone	0.079320	1
triamcinolone	0.074891	1
prednisolone	0.072578	1

ind_id	alcohol withdrawal delirium	previous reco
chlordiazepoxide	1.000000	1
oxazepam	1.000000	1
diazepam	1.000000	1
clorazepate	0.554800	0
etizolam	0.442906	0
halazepam	0.442906	0
chlormezanone	0.442906	0
ethyl loflazepate	0.442906	0
meprobamate	0.442906	0
trifluoperazine	0.442906	0

ind_id	hypersomatotropic gigantism	previous reco
lanreotide	1.000000	1
pegvisomant	0.772644	0
lisuride	0.381525	0
bromocriptine	0.211878	0
octreotide	0.132879	0
abacavir	0.000000	0
olsalazine	0.000000	0
oritavancin	0.000000	0
orciprenaline	0.000000	0
oprelvekin	0.000000	0

Figure 3: TOP 10 drug recommendations for Abdominal Actinomycosis, Addison's disease, Alcohol withdrawal and Gigantism. The second columns shows the known drug-disease associations

We notice in the first table that the drugs with top scores are penicilin-based compounds, like the only known and recommended drug for this disease.

In the second table, we can see two interesting things: some drugs that were not known are better ranked that some known drug associations, and the Dalfampridine recommendation makes sense. Indeed Dalfampridine is usually used to improve walking in people with multiple sclerosis and is a potassium channel blocker; and Addison's disease is associated to Hyperkalemia (which means high level of potassium). Additionally, Addison's disease is known to be caused by a deficience of cortisone and aldosterone; and some studies link those hormones to the Multiple Sclerosis (see [16]).

In the third table, Clorazepate is the first prediction of not already known drugs in our data. It is however commonly prescribed, and our method could fond it again.

In the fourth table, Pegvisomant is highly recommended. What is interesting is that even if it is not currently prescribed for Gigantism, some papers recommended it as a treatment (see [17, 18]).

5 Conclusion

We implemented a new method of drug repurposing based on Alaimo's recommendation method (DT-Hybrid, see [14]). We used the Statistically Significant Connectivity Map metric (ssCMap) to compute the similarities between drugs based on their gene signatures, and between diseases based on their gene phenotypes. One could use other metrics, like XCos, to see if it could improve our results.

By using cross-validation, we chose the best parameters to use in the computation of the method according to their ability to "recover" deleted drug-disease associations.

5.1 Interpretation of the results

We must note that the result of 61.54% and 74.56% for the recovery metrics (see 4.1) is quite good considering the form of the data. Indeed, it is important to note that the random partitioning method associated with the cross-validation can cause the isolation of some nodes in the network on which the tests are being performed. A main limitation of recommendation algorithms like in our method is the inability to predict new interactions for drugs or diseases for which no information is available. This implies that in the presence of isolated nodes a bias is introduced in the evaluation of results [13].

Furthermore, we computed recommendations for all diseases and drugs even if we had no information on them in the signature or association tables by filling the tables with 0. As a consequence, bad final scores in the recommendation tables might be caused by a lack of information rather than an incompatibility between the drug and disease. We must note that it is preferable, since drug repurposing should rely on "worst-case" scenario in the face of uncertainty as a safety precaution.

Additionally, it would be interesting to check for all of the recommended drugs if they were reported as effective for the corresponding diseases in the literature. Unfortunately, we did not have the names associated to most of the diseases (only ids like 'c3665869 ').

References

[1] Subramanian et al. (November 2017). *A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles*

[2] Q. Duan et al. (August 2016). *L1000CDS^2: LINCS L1000 characteristic direction signatures search engine*

[3] Clark NR, Hu KS, Feldmann AS et al. (2014). *The characteristic direction: a geometrical approach to identify differentially expressed genes.*

[4] Musa et al. (2018). *A review of connectivity map and computational approaches in pharmacogenomics*

[5] Zhang SD, Gant T. A (2008). *A simple and robust method for connecting small-molecule drugs using gene-expression signatures.*

[6] J. Cheng et al. (2013). *Evaluation of analytical methods for connectivity map data.*

[7] J. Cheng et al. (2014). *Systematic evaluation of connectivity map for disease indications.*

[8] J. Lamb et al. (2006). *The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease.*

[9] Qu XA, RajPal DK. (2012). *Applications of connectivity map in drug discovery and development.*

[10] F. Ioro et al. (2010). *Discovery of drug mode of action and drug repositioning from transcriptional responses.*

[11] F. Ioro et al. (2013). *Network based elucidation of drug response: from modulators to target.*

[12] C. Pacini, F. Ioro et al. (2013). *DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data*

[13] S. Alaimo, R. Giugno, A. Pulvirenti et al. (2016). *Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning.*

[14] S. Alaimo, A. Pulvirenti, R. Giugno et al. (2013). *Drug–target interaction prediction through domain-tuned network-based inference.*

[15] S. Alaimo, V. Bonnici, D. Cancemi et al. (2015). *Dt-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference.*

[16] J. Moynihan and H. Moore (2010). *Endocrine system dynamics and MS epidemiology.*

[17] Müssig K, Gallwitz B et al. (2007). *Pegvisomant treatment in gigantism caused by a growth hormone-secreting giant pituitary adenoma.*

[18] Naila Goldenberg, Michael S. Racine et al. (2008). *Treatment of Pituitary Gigantism with the Growth Hormone Receptor Antagonist Pegvisomant.*