JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009

Pp. 241–251
DOI: 10.1089/cmb.2008.10TT

Identifying Network of Drug Mode of Action by Gene Expression Profiling

FRANCESCO IORIO,[1,2] ROBERTO TAGLIAFERRI,[2] and DIEGO DI BERNARDO[1,3]

ABSTRACT

Drug mode of action (MOA) of novel compounds has been predicted using phenotypic features or, more recently, comparing side effect similarities. Attempts to use gene expression data in mammalian systems have so far met limited success. Here, we built a drug similarity network starting from a public reference dataset containing genome-wide gene expression profiles (GEPs) following treatments with more than a thousand compounds. In this network, drugs sharing a subset of molecular targets are connected by an edge or lie in the same community. Our approach is based on a novel similarity distance between two compounds. The distance is computed by combining GEPs via an original rank-aggregation method, followed by a gene set enrichment analysis (GSEA) to compute similarity between pair of drugs. The network is obtained by considering each compound as a node, and adding an edge between two compounds if their similarity distance is below a given significance threshold. We show that, despite the complexity and the variety of the experimental conditions, our approach is able to identify similarities in drug mode of action from GEPs. Our approach can also be used for the identification of the MOA of new compounds.

Key words: connectivity map, drug mode of action, gene set enrichment analysis, ranks merging, similarity networks.

1. INTRODUCTION

IDENTIFYING PATHWAYS THAT MEDIATE a drug mode of action (MOA) is a key challenge in biomedicine (Ambesi-Impiombato and di Bernardo, 2006; di Bernardo et al., 2005; Staunton et al., 2001; Terstappen et al., 2007; Yao and Rzhetsky, 2008). Several chemo-informatics tools to analyze chemical similarities between small-molecules are available (Medina-Franco et al., 2007; Miller, 2002; Rhodes et al., 2007). Recently, text-mining has been proposed as an approach to estimate drug MOA similarities using similarity in drug clinical side-effects (Campillos et al., 2008; Yeh et al., 2006). All of these methods require extensive prior knowledge on the compounds (e.g., molecular structure, side effects) or have been tested only on simple organisms (di Bernardo et al., 2005).

[1]Systems and Synthetic Biology Lab, TeleThon Institute of Genetics and Medicine (TIGEM), Naples, Italy.
[2]Neural and Robotic Networks (NeuRoNe) Lab, Department of Mathematics and Computer Science, University of Salerno, Fisciano, Salerno, Italy.
[3]Department of Computer Science and Systems, University "Federico II" of Naples, Naples, Italy.

Here we developed an approach to draw a "drug network" from gene expression profiles (GEPs) following drug treatments in human cell lines. In the network, drugs with similar MOA are connected or lie in the same "community."

We obtained the network by analyzing a public dataset of genome-wide GEPs, the *Connectivity Map* (cMap), consisting of expression profiles from five different human cell lines treated with 1309 different compounds, at different concentrations (Lamb, 2007; Lamb et al., 2006).

The cMap consists of a set of experiments (batches). Each batch is composed of two or more microarray hybridizations of a compound-treated cell line and one or more hybridizations of the untreated cell line (as negative control). The number of treatments and controls per batch can vary in the number of total treatments across batches per single drug. Within a batch, the change in gene expression in a cell line after each treatment with a given compound is computed considering the differential gene expression values of the treated cell versus the untreated one (or the set of untreated ones). Each treatment with a compound in a batch thus yields a genome-wide differential GEP.

In Lamb et al. (2006), the authors provide a web tool to find connections between a well-defined subset of microarray probe identifiers (i.e., a *signature* made of gene transcripts) and the GEPs of the cMap. A ranked list of microarray probe identifiers is obtained by sorting each GEP according to its differential expression values (from the most up-regulated gene to the most down-regulated one). The tool makes use of a modified version of the Gene Set Enrichment Analysis (GSEA) method in which the signature is compared to each ranked list (GEP) to determine whether up-regulated signature genes tend to appear at the top of the list and down-regulated signature genes at the bottom (Subramanian et al., 2005). In Lamb et al. (2006), the signature is chosen from the literature by selecting genes known to be involved in a particular process of interest. The signature is then compared to each GEP in the cMAP dataset in order to identify drugs that are likely to act on the same process. Inspired by this work, we developed a novel approach and applied it to the cMap dataset to construct a drug similarity network.

2. RESULTS

We developed an approach able to automatically select an *optimal signature* for each drug in the cMap dataset and to quantify similarities among the MOA of all the compounds in the dataset. From the similarity measure, we then built a network where each compound is a node and two compounds are connected by an edge if their similarity is below a given significance threshold. We then assessed the usefulness of the resulting network by verifying if the known similarity in MOA among drugs corresponded to connected nodes in the network.

In our approach, we merged ranked lists (GEPs) obtained from treatments with the same drug on different cell lines and for different concentrations, to obtain a *prototype ranked list* (PRL). The PRL is computed by means of a novel rank aggregation method combining the *Kruskal Algorithm* (for the computation of the *Minimum Spanning Tree* in graphs), the *Borda Rank Aggregation Method*, and the *Spearman Foot-Rule* distance (to compute the agreement between ranked lists) (Cormen et al., 1990; Diaconis and Graham, 1977; Parker, 1995). From each of these PRLs, an *optimal signature* of genes is computed for each drug by selecting the top 250 genes and the bottom 250 ones. The optimal signature summarizes the effect of a drug, and it is independent of cell line and concentration effects. We assumed that the more the optimal signature of a drug is enriched in a PRL of another drug (according to GSEA), the more the two drugs will share a similar MOA. We quantified this similarity using a novel metric based on the GSEA in order to yield a similarity distance (for details, see Section 4, Methods).

2.1. Similarity metric assessment

The empirical *probability density function* (pdf) of the similarity distance among all the 1309 drugs is shown in Figure 1 (black curve). The total number of similarity distances is $(1309^2 - 1309)/2 = 856{,}086$ (since there are treatments with 1309 drugs in the dataset and our similarity distance is symmetric). In Figure 1, the gray points represent the pair-wise similarity distances among drugs contained in five

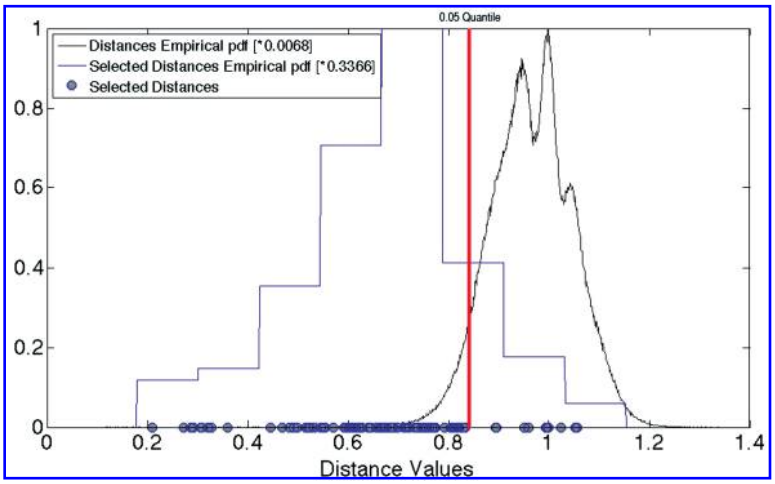

FIG. 1. Empirical probability density functions of the drug similarity distances among all 1309 drugs (black line) and among the drugs within the six test groups (blue line). Gray points represent the similarity distances for the six test groups; the red line marks the 5% quantile for the empirical probability density function (black line).

predefined test groups known to have similar MOA. The composition of each test group is shown in Table 1a: (1) *histone deacetylase inhibitors*; (2) *COX2 inhibitors*; (3) *antipsychotics*; (4) *heat shock protein 90 (Hsp90) inhibitors*; and (5) *anti-diabetics*. Note that similarity distances in Figure 1 (gray points) were computed only among drugs within the same group, for a total of 95 values. Drugs with similar MOA tend to have small values of similarity distance when compared to the distribution of the similarity distances across the whole population of 1309 drugs. In order to quantify the statistical significance of this observation, we noted that only 5% of distances across the whole population of 1309 drugs were less than 0.8339 (the red line in Fig. 1). On the contrary, 90% of the subset of 95 distances computed among drugs in the five test groups was less than 0.8339. The difference between these two percentages was not obtained by chance. (The χ^2 test p-value against the null hypothesis that the difference between these two percentages is random is $\approx 10^{-16}$).

To further confirm this finding, we performed a *permutation test* by considering 10,000 subsets of 95 distances picked at random from the set of all the possible 856,086 distances in the empirical distribution. We computed the average distance within each of the 10,000 subsets. We then computed the difference between the average distance within the whole population and average distance within each subset. The larger this difference, the more similar are the compounds in a subset relative to the whole population. We observed that the difference computed, using as a subset the 95 distances in our test groups, had the largest value among all the 10,000 differences obtained with the random subsets. Thus, the empirical p-value against the null hypothesis where the labeling does not influence the average distances is $p_{perm} < 10^{-4}$.

We assessed the reliability of our similarity distance by labeling all the compounds in the cMAP according to their Anatomical Therapeutic Chemical (ATC) classification code (Schwabe, 1995), which classifies drugs according to their therapeutic and chemical characteristics. Since only 768 out of the 1309 compounds have an ATC code, we performed our analysis on the subset of $\binom{768}{2} = 294{,}528$ similarity distances out of the total set of distances. We then ranked distances in this subset (i.e., edges in the network) in ascending order, and computed the Receiver Operator Characteristic (ROC) as shown in Figure 2 and Table S2. (See online Supplementary Materials at *www.liebertonline.com*.) The ROC is computed by considering as positive predictions the first k smallest similarity distances in the ranked list, with $k = 1, \ldots, 280{,}875$. We considered as a *true positive prediction* (TP) a similarity distance between two drugs which share the same ATC prefix. For each k, we computed the *Positive Predicted Value* (PPV), by computing the percentage of true positives (TP) out of the k predictions. We also computed the expected PPV had the similarity distances been chosen at random (dashed line in Fig. 2). The results show that our approach is able to infer meaningful interactions among drugs.

TABLE 1. DRUG TEST-GROUPS, IDENTIFIED COMMUNITIES AND DRUG NEIGHBORHOODS

a - Drug Test-Groups used to assess similarity distances

Histone Deacetylase Inhibitors	
trichostatin A, HC toxin, rifabutin, vorinostat, scriptaid, valproic acid,	
COX2 Inhibitors	
LM-1685, celecoxib, rofecoxib, SC-58125	
Antipsychotics	
thioridazine, trifluoperazine, fluphenazine, perphenazine, chlorpromazine, chlorpropamide, prochlorperazine, haloperidol	
Heat Shock Protein 90 Inhibitors	
geldanamycin, monorden, alvespimycin, puromycin, celastrol, tanespimycin, MG-132, MG-262, thiostrepton, anisomycin	
Antidiabetics	
troglitazone, rosiglitazone	

b - Some Drug-Communities identified in the network obtained by fixing the distance threshold level to 0.6

Community n.1		Distance Average
Cardiac Glicosyde (All founded in Digitalis and Sthrophantum) and two Protein Synthesis inhibitors. All of them (but Strophanthidin) inhibits Caspase-3	anisomycin cycloheximide digitoxigenin digoxigenin digoxin [C01AA05] helveticoside lanatoside_C [C01AA06] ouabain proscillaridin [C01AB51] strophanthidin	0.35
Community n.2		**Distance Average**
Histone Deacetylase Inhibitors	HC_toxin MS-275 rifabutin scriptaid vorinostat	0.42
Community n.3		**Distance Average**
benzimidazoles	nocodazole mebendazole	0.45
Community n.4		**Distance Average**
Heat Shock Protein 90 Inhibitors	alvespimycin geldanamycin monorden tanespimycin	0.47
Community n.5		**Distance Average**
Anthracyclines, anthracenediones, CDK inhibitors and a topo I inhibitor	GW-8510 H-7 alsterpaullone camptothecin daunorubicin doxorubicin [L01DB01] mitoxantrone tyrphostin_AG-825	0.48
Community n.6		**Distance Average**
Aantihistamines, anticholinergics	astemizole [R06AX11] mefloquine suloctidil	0.48
Community n.7		**Distance Average**
Flavones, Flavonoids and an alkalaoid from plants of similar species. All of them modulate PPAR-Gamma (that is known to protect againts diabetes).	apigenin chrysin harmine luteolin	0.5
Community n.8		**Distance Average**
Proteasome inhibitors and anticancers	MG-132 MG-262 celastrol lomustine parthenolide phenoxybenzamine piperlongumine puromycin thiostrepton withaferin_A	0.54
Community n.9		**Distance Average**
two antipsychotics (phenotiazines) and an opioid receptor agonist	loperamide perphenazine trifluoperazine	0.56
Community n.10		**Distance Average**
PI3K inhibitors	sirolimus wortmannin	0.56
Community n.11		**Distance Average**
quaternary ammonium compounds, disinfectants and antiseptics	benzethonium_chloride methylbenzethonium_chloride	0.56
Community n.12		**Distance Average**
estrogen and progesterone receptors inhibitors	exisulind sulindac sulfide	0.57
Community n.13		**Distance Average**
ergot alkaloids derivatives	bromocriptine methylergometrine	0.58

c - Assessment of the network obtained by fixing the distance threshold level to 0.8

Testing Drug	Neighbors	Distance
trichostatin A	rifabutin [J04AB04]	0.6843
	vorinostat	0.7331
	HC Toxin	0.7547
	scriptaid	0.761
	MS-275	0.7624
	Prestwick-1080	0.788
Testing Drug	**Neighbors**	**Distance**
perphenazine [N05AB03]	**trifluoperazine [N05AB06]**	0.5207
	astemizole [R06AX11]	0.567
	loperamide [A07DA03]	0.5698
	terfenadine [R06AX12]	0.6085
	niclosamide	0.6121
	metergoline [G02CB05]	0.6223
	mefloquine [P01BC02]	0.6229
	methylbenzethon	0.636
	alexidine	0.6405
	gossypol	0.6642
	tetrandrine	0.6654
	fluphenazine [N05AB02]	0.6697
	maprotiline	0.6709
	mometasone	0.6715
	chlorprothixene	0.6795
	levomepromazine [N05AA02]	0.6838
	thapsigargin	0.6848
	haloperidol [N05AD01]	0.6853
	metitepine	0.6866
	metixene [N04AA03]	0.6873
	prenylamine [C01DX02]	0.6891
	desipramine [N06AA01]	0.6965
	calmidazolium	0.6966
	zuclopenthixol [N05AF05]	0.6984
	prochlorperazine [N05AB04]	0.6989
Testing Drug	**Neighbors**	**Distance**
rofecoxib [M01AH02]	**SC-58125**	0.6455
	dexverapamil	0.6914
	exemestane [L02BG06]	0.7106
	LM-1685	0.7464
	chlorzoxazone [M03BB03]	0.7758
	lobeline	0.7984
Testing Drug	**Neighbors**	**Distance**
alpha-estradiol	**genistein**	0.8
	exemestane	0.79
Testing Drug	**Neighbors**	**Distance**
valproic acid [N03AG01]	rifabutin [J04AB04]	0.6844
	HC_toxin	0.7108
	vorinostat	0.7337
	scriptaid	0.773
	acetylsalicylic_acid [N02BA01]	0.7865
Testing Drug	**Neighbors**	**Distance**
verapamil [C08DA01]	**dexverapamil**	0.75

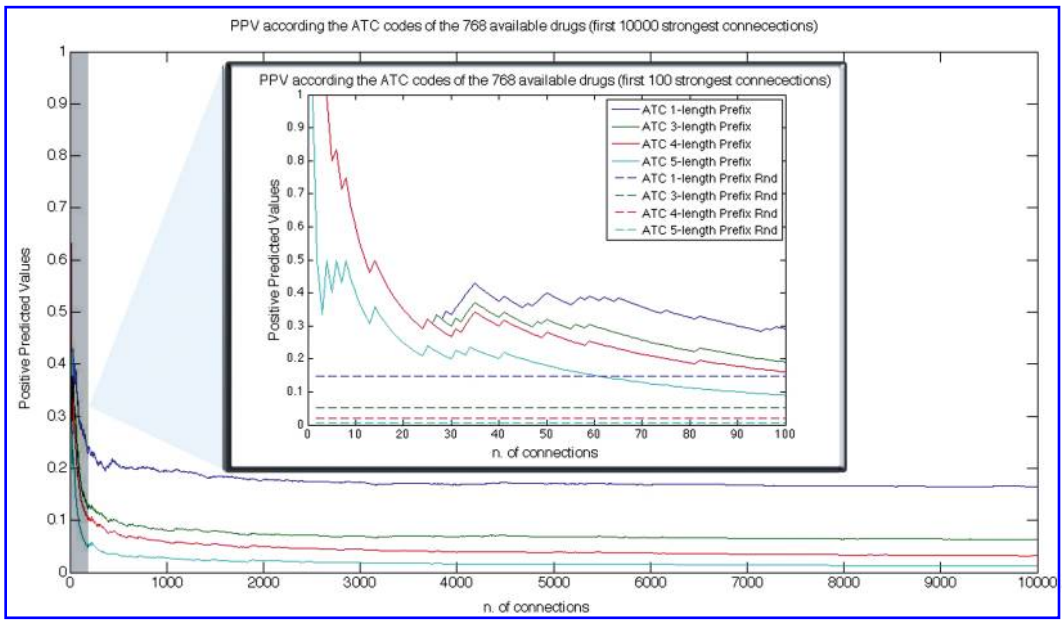

FIG. 2. Performance of the drug similarity distance in identifying compound mode of action (MOA) according to the Anatomical Therapeutic Chemical (ATC) classification. Positive predicted value (PPV) versus number of connections ranked according to the similarity distance (from the smallest to the largest). A true positive prediction is defined as two drugs sharing the ATC code prefix.

2.2. *Analysis of the drug network*

Figure 3 shows networks where each node corresponds to a compound, and two nodes are connected by an edge, if the corresponding similarity distance is less than a predefined threshold (0.2, 0.3, and 0.4 for networks A, B, and C, respectively). The width of an edge is inversely proportional to the distance between the drugs connected by the edge.

In network A, there are two connected components: one consists of *cardiac glycosides* (as also shown by their ATC code): *digoxigenin*, *digoxin* [C01AA05], *digitoxigenin* (*steroids* found in some species of *Digitalis*, cardiac glycosides); *ouabain* (*endogenous hormone* found in the ripe seeds of the African plant *Strophantus*); *proscillaridin* [C01AB51] (a *cardiac glycoside* from the plant *Scilla Maritima*); *lanatoside C* [C01AA06]; and *helveticoside* (two other *cardiac glycoside*).

The second component consists of *protein synthesis inhibitors: cephaeline* and *emetine* [P01AX02], two *ipecac alkaloids* (Arany et al., 2008). This shows the predictive ability of the network since similarity distances are computed using only the GEPs without knowledge of the MOA of the drugs.

Increasing the threshold level to 0.3, the network grew in a consistent and coherent way (Fig. 3B): new connections appeared in the *cardiac glycosides component*, whereas *cycloheximide* joins the protein synthesis inhibitors; a cluster of *Histone Deacetylase Inhibitors* (*vorinostat*, *scriptaid*, *HC toxin*), containing also the bactericidal antibiotic *rifabutin* [J04AB04], appeared, and two *inhibitors of the Hsp90 chaperone* (*geldanamycin* and *alvespymicin*) were linked together.

Further increasing the threshold to 0.4 yielded the network in Figure 3C. Several drugs were added to the network: *monorden* joined the other two Hsp90 inhibitors in agreement with its known MOA; a small cluster containing *cyclic-dependent kinases inhibitors* (*doxorubicin* [L01DB01], *GW-8510*, and *alsterpaullone*) and *H-7* appeared, *luteolin* and *apigenin* (a *flavone* and a *flavonoid*) become connected, as well as *campthotecin* and *irinotecan* [L01XX19] (*DNA enzyme topoisomerase I inhibitors*), *astemizole* [R06AX11] and *terfenadine* [R06AX12] (two *antihistamines*), *MG-262* and *MG-132* (two *proteasome inhibitors*), *dexverapamil* and *examestane* (two *L-type calcium channel blockers*). Additionally, four pairs of drugs (whose similarity in MOA is not well confirmed) were linked together. A weak edge between *lanatoside-C* and *anisomycin* connected the clusters of cardiac glycosides to protein synthesis inhibitors. This could be due to the fact that these two drugs share the inhibition of *caspase-3* (Piccioni et al., 2004).

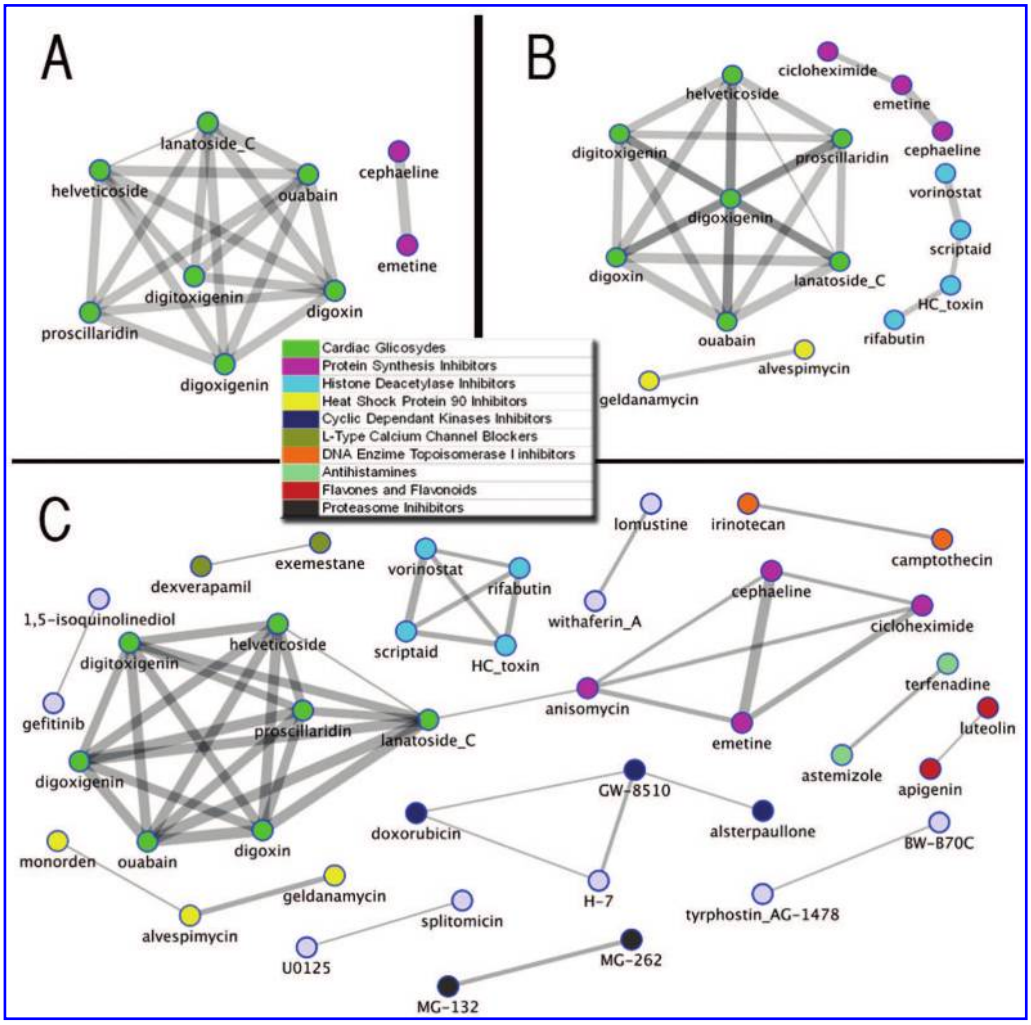

FIG. 3. Drug networks obtained by select threshold level for the similarity distance to 0.2 **(A)**, 0.3 **(B)**, and 0.4 **(C)**, respectively. Each node represents a compound, and its color represents its known mode of action (MOA). Two nodes are linked by an edge if their similarity distance is below the predefined threshold.

Increasing the distance threshold level to 0.6 yields a network of 158 drugs with 379 edges. In order to assess it, we identified its *communities*. In graph theory, a community is a set of nodes connected to each other by a large number of edges, but with few edges connecting it to the rest of the network. We used a modified version of the Girvan-Newman algorithm to identify communities (Girvan and Newman, 2002). A view of the network, highlighting the 33 identified communities, is shown in Figure 4. The composition of some of the communities is provided in Table 1b.

The network obtained by increasing the similarity distance threshold to 0.8 included 1252 drugs with 18,030 edges. The network has 240 communities containing an average number of 2.89 drugs (the largest one contains 13 nodes). To assess this network, we selected six drugs: (1) *trichostatin A* (a *histone deacetylase inhibitor*); (2) *rofecoxib* [M01AH02] (a *COX2 inhibitor*); (3) *perphenazine* [N05AB03] (an *antipsychotic* drug of the *phenothiazine group*); (4) *alpha-estradiol* (an *estrogen receptors modulator*); (5) *valproic acid* [N03AG01] (a *histone deacetylase inhibitor*); and (6) *verapamil* [C08DA01] (an *L-type calcium channel blocker*). The analysis of the neighbors of these drugs in the network is summarized in Table 1c.

In order to compare our results with those of the cMap online query tool and in order to assess how much our optimal signatures are useful to query the cMap in the standard way, we used the optimal signatures obtained with our method for the above six drugs. Results of this comparison are summarized in Table S1. (See online Supplementary Materials at *www. liebertonline.com*.) The standard query tool achieved a good

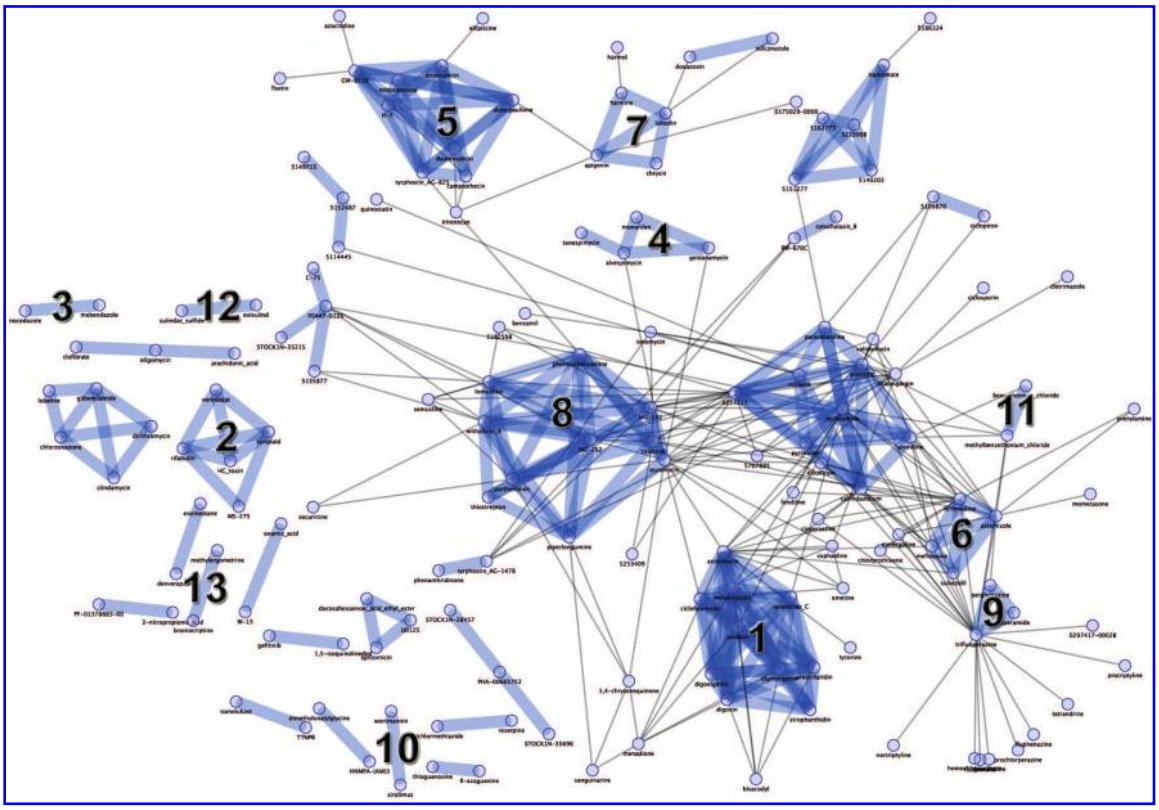

FIG. 4. Drug network obtained selecting a threshold of 0.6 for the similarity distance. Each node represents a compound. Two nodes are linked by an edge if their similarity distance is below the predefined threshold. The network communities identified with the Girvan-Newman algorithm are highlighted and reported in Table 1a.

performance in selecting drugs with similar MOA if, as a query signature, we used the optimal signature obtained with our approach. All the used drug signatures and the results obtained with the cMap tool are present in the Supplementary Materials. (See online Supplementary Materials at *www. liebertonline.com*.)

2.3. Robustness of the drug network

In order to assess the robustness of our approach, we computed drug similarity distances using optimal signatures of 50 and 150 genes. All of the obtained networks are given in the Cytoscape© format in the Supplementary Materials. (See online Supplementary Materials at *www. liebertonline.com*.)

We observed that, by using the same cutoff threshold value with queries of different lengths, the network obtained with the smallest query size always contained, as subnetwork, the networks obtained with the larger query sizes. This means that, as the query size increases, our similarity measure grows in specificity; however, the overall structure of the network does not change.

3. DISCUSSION

In this work, we showed that, starting from a compendium of genome-wide GEPs following treatments with different drugs, it is possible to obtain a drug MOA similarity network capable of identifying drugs that share their molecular targets. This method also permits the classification of a novel compound by simply computing its prototype ranked list, obtained by treating a set of cell lines with the compound, and then its optimal signature. All that is needed to accomplish this classification is a set of genome-wide microarray hybridizations following treatments with the novel compound (on a sufficiently large variety of cell cultures, with at least an untreated hybridization per cell line, as negative control). From these, the prototype ranked list of the novel compound can be estimated (see Section 4, Methods) and the compound

can be integrated in the network. Normalization procedures are not needed, and the existing network does not have to be modified when a novel node is added to it. By analyzing the structure of the network, it is possible to formulate hypotheses on the MOA of the novel compound, by checking to which other drugs in the network it is connected.

Our approach represents a powerful method able to identify MOA using only gene expression data that could help in the complex drug discovery process.

4. METHODS

4.1. Prototype list generation

We assembled a PRL for each drug by aggregating all the ranked lists of differentially expressed genes (GEPs) in the cMAP dataset obtained by treating cells with the same drug on different cell lines and with different concentrations. In our aggregation method, we made use of the following tools: a measure of the distance between two ranked lists (Spearman's Foot-Rule), a method to merge two or more ranked lists (the Borda Merging Method) and an algorithm to obtain a single ranked list from a set of them in a hierarchical way (the Kruskal Algorithm) (Cormen et al., 1990; Diaconis and Graham, 1977; Parker, 1995).

Similarly to a hierarchical clustering method, our algorithm first computes the pair-wise Spearman's Foot-Rule distances between all the ranked lists obtained with the same drug. Then it merges the two lists that are closest to each other according to this distance (with the Borda Merging Method), obtaining a new ranked list. The new list is used to replace the two lists that have been merged together, and the Spearman's Foot-Rule distances are recomputed. This procedure is repeated until only one ranked list remains. A detailed description of our merging algorithm is provided in the Supplementary Materials. (See online Supplementary Materials at *www. liebertonline.com.*)

An example of the application of our method is shown in Figure 5. We start from the pair-wise Spearman's Foot-Rule distances among all the ranked lists obtained treating cells with *Alvespimycin*, an inhibitor of the Hsp90. Each node of the tree is a ranked list of genes (GEP), and the Euclidean distances between the nodes in the Figure 5 reflect the Spearman's Foot-Rule distances between the ranked lists that

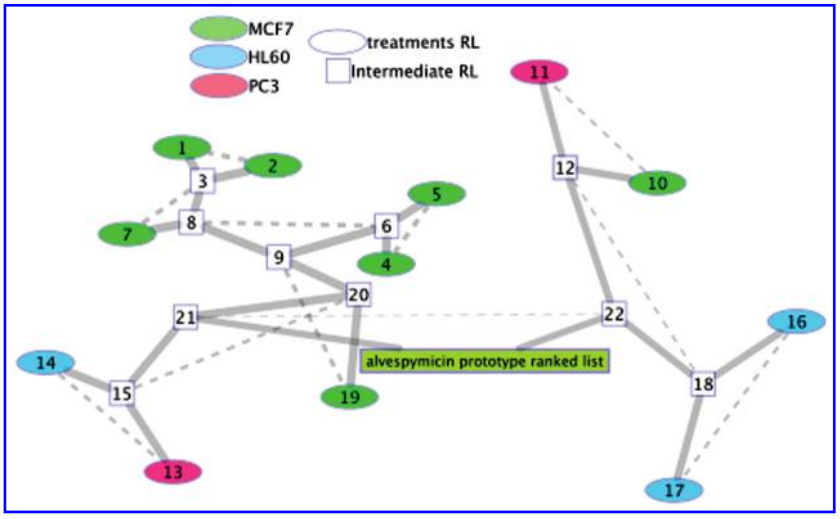

FIG. 5. Example of the computation of the prototype ranked list (PRL) for a compound. Each leaf (ellipses) represents a ranked list obtained by treating a cell line with *Alvespymicin*. Colors specify the treated cell line. Each internal node (square) represents an intermediate ranked lists obtained by merging the two lists represented by its children nodes. The width of an edge connecting two nodes is inversely proportional to the Spearman's Foot-Rule distance between the ranked lists represented by those nodes. The root of the tree is the final PRL for *Alvespimycin* (large green rectangle). Solid lines indicate childhood relationships, while nodes connected by a dashed line indicate siblings.

they represent. The first two lists that the algorithm merges are those corresponding to nodes 1 and 2 (the closest ones). These two lists are merged with the Borda Merging Method, yielding the new list represented by node 3. The algorithm continues by merging nodes 4 and 5 (the second closest pair in the set), yielding the list in node 6. This process iteratively continues until the lists represented by nodes 21 and 22 are the only present ones. They are merged, and the *Alvespymicin* PRL is obtained.

If applied to the cMap, this approach is able to correctly merge ranked lists of differentially expressed genes obtained by multiple treatments with the same drug. Namely, there are cases in which several treatments with a single drug on a particular cell line are available, but only one treatment with the same drug is present on a different cell line. In such a case, applying a single majority voting method (i.e., the Borda Merging Method) will lead to a final merged list in which the response of the cell line with a single treatment is not represented at all.

It is possible that, with our merging procedure, an outlier (i.e., a systematic error in the hybridization leading to a wrong ranked list) could be overweighed. This can be avoided by adding a pre-filtering step to remove wrong data. However, our results show that our method is robust when applied to the cMap dataset.

4.2. Drug distances

Once a PRL has been obtained for each drug $\boldsymbol{d}$ in the dataset, we created an *optimal signature* $\{p, q\}$. To this end, we selected the top-ranked 250 genes of each PRL and the bottom-ranked 250 ones (p and q, respectively). We considered this signature of genes as a general cellular response to the drug. In other words, we isolated sets of genes that seemed to vary in response to the drug across different experimental conditions (e.g., different cell lines, different dosages).

Now, given the optimal signature of drug $\boldsymbol{d}$, $p = \{p_1, \ldots, p_n\}$ (up-regulated) and $q = \{q_1, \ldots, q_m\}$ (down-regulated), we defined as the distance between drug $\boldsymbol{d}$ and drug $\boldsymbol{x}$ the *Inverse Total Enrichment Score* (TES) of the drug $\boldsymbol{d}$ signature $\{p, q\}$, with respect to the PRL of drug $\boldsymbol{x}$, as follows:

$$TES_{d,x} = 1 - \frac{ES_x^p - ES_x^q}{2}.$$

Here, $ES_x^r (r \in \{p, q\})$ is the Enrichment-Score of the optimal signature with respect to the PRL of $\boldsymbol{x}$. ES_x^r ranges in $[-1, 1]$, it is a measure based on the Kolmogorov-Smirnov statistics, and it quantifies how much a set of genes is at the top of a ranked list (Subramanian et al., 2005). The closer that this measure is to 1, the more the genes are at the top of the list, whereas the closer to -1, the more the genes are at the bottom of the list. $TES_{d,x}$ ranges in $[0, 2]$, it takes into account two sets of genes, and it checks how much the genes in the first set (p) are placed at the top of the $\boldsymbol{x}$ PRL and how much the genes in the second set (q) are placed at the bottom. The more these two statements are true, the more the value of $TES_{d,x}$is close to 0.

Once we obtained an optimal signature $\{p, q\}$, for each drug $\boldsymbol{d}$, we computed a vector T_d containing the TES of the $\boldsymbol{d}$ optimal signature with respect to all the PRLs.

Finally, we grouped all these line vectors in a single matrix:

$$M = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \end{bmatrix}.$$

Note that this is a 1309×1309 matrix, whose entries range in $[0, 2]$, and it has exactly one entry equal zero per row and per column. Therefore, with a proper permutation of the rows, we can place all these zeros entries on the diagonal. To obtain the final distance matrix, we simply provided symmetry to M (its permuted version):

$$M_{ES} = \frac{M + M^T}{2}.$$

The entry i, j (or j, i) of this matrix contains the distance between drugs i and j.

4.3. Community identification

In order to identify communities in our drug network, we implemented (in Matlab©) the *Girvan-Newman* algorithm for the identification of communities of nodes in a network (Girvan and Newman, 2002). A network is said to have a *Community Structure* if its set of nodes can be partitioned into groups such that the density of connections (in terms of number of edges) within the groups is higher than the density among groups. The Girvan-Newman algorithm proceeds as follows: (1) it computes the shortest paths (in terms of number of edges) between each couple of nodes; (2) for each edge, it counts the number of shortest paths that pass through it (the *betweenness* or the *centrality* of the edge); and (3) it removes the edge with the highest betweenness from the network.

These steps are iteratively repeated until no edges remain in the network. Taking into account the components of the network that eventually disconnected after an edge removing, this procedure gives a hierarchy of communities in output.

We implemented this algorithm with the additional following modification: (1) we used the weighted shortest path between nodes (a path between a pair of nodes is the shortest one if the sum of the weights on the edges composing it is minimal); (2) if a community becomes a *singleton* (i.e., it contains one only node), then it is removed from the network; and (3) if the total number of disconnected components decreases, then the computation is stopped.

ACKNOWLEDGMENTS

We wish to thank Luisa Cutillo and Francesco Napolitano for a number of insightful discussions. All the figures containing networks have been obtained using Cytoscape©. This work was supported by a grant of Associazione Italiana Ricerca Cancro ("Inferring Gene Networks and Compound Mode of Action by Expression Profiling") to D.d.B. and to a grant of Fondazione TeleThon to D.d.B.

DISCLOSURE STATEMENT

No conflicts of interest are present.

REFERENCES

Ambesi-Impiombato, A., and di Bernardo, D. 2006. Computational biology and drug discovery: from single-target to network drugs. *Curr. Bioinform.* 1, 3–13.

Arany, Z., Wagner, B.K., Ma, Y., et al. 2008. Gene expression-based screening identifies microtubule inhibitors as inducers of PGC-1α and oxidative phosphorylation. *Proc. Natl. Acad. Sci. USA* 105, 4721–4726.

Campillos, M., Kuhn, M., Gavin, A.C., et al. 2008. Drug target identification using side-effect similarity. *Science* 321, 263–266.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L. 1990. Minimum spanning trees. In: *Introduction to Algorithms*. MIT Press, Cambridge, MA.

di Bernardo, D., Thompson, M.J., Gardner, T.S., et al. 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383.

Diaconis, P., and Graham, R. 1977. Spearman's foot-rule as a measure of disarray. *J. R. Statist. Soc.* 39, 262–268.

Girvan, M., and Newman, M.E. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 7821–7826.

Lamb, J. 2007. The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* 7, 54–60.

Lamb, J., Crawford, E.D., Peck, D., et al. 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935.

Medina-Franco, J.L., Maggiora, G.M., Giulianotti, M.A., et al. 2007. A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Des.* 70, 393–412.

Miller, M.A. 2002. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* 1, 220–227.

Parker, J.R. 1995. Voting methods for multiple autonomous agents. *Proc. 3rd Aust. N.Z. Conf. Intell. Inform. Syst. (ANZIIS-95).*

Piccioni, F., Roman, B.R., Fischbeck, K.H., et al. 2004. A screen for drugs that protect against the cytotoxicity of polyglutamine-expanded androgen receptor. *Hum. Mol. Genet.* 13, 437–446.

Rhodes, J., Boyer, S., Kreulen, J., et al. 2007. Mining patents using molecular similarity search. *Pac. Symp. Biocomput.* 304–315.

Schwabe, U. 1995. *ATC Code*. Wissenschaftliches Institut der AOK, Bonn, Germany.

Staunton, J.E., Slonim, D.K., Coller, H.A., et al. 2001. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* 98, 10787–10792.

Subramanian, A., Tamayo, P., Mootha, V.K., et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550.

Terstappen, G.C., Schlupen, C., Raggiaschi, R., et al. 2007. Target deconvolution strategies in drug discovery. *Nat. Rev. Drug Discov.* 6, 891–903.

Yao, L., and Rzhetsky, A. 2008. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res.* 18, 206–213.

Yeh, P., Tschumi, A.I., and Kishony, R. 2006. Functional classification of drugs by properties of their pairwise interactions. *Nat. Genet.* 38, 489–494.

Address reprint requests to:
Dr. Diego di Bernardo
TeleThon Institute of Genetics and Medicine (TIGEM)
Via Pietro Castellino 111
80131 Naples, Italy

E-mail: dibernardo@tigem.it

This article has been cited by:

1. Peter Csermely, Tamás Korcsmáros, Huba J.M. Kiss, Gábor London, Ruth Nussinov. 2013. Structure and dynamics of molecular networks: A novel paradigm of drug discovery. *Pharmacology & Therapeutics* . [CrossRef]
2. Haisu Ma, Hongyu Zhao. 2013. Drug target inference through pathway analysis of genomics data. *Advanced Drug Delivery Reviews* . [CrossRef]
3. Xiaoyan A. Qu, Deepak K. Rajpal. 2012. Applications of Connectivity Map in drug discovery and development. *Drug Discovery Today* **17**:23-24, 1289-1298. [CrossRef]
4. H. Ma, H. Zhao. 2012. FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. *Bioinformatics* **28**:20, 2662-2670. [CrossRef]
5. Li Xie, Sarah L. Kinnings, Lei Xie, Philip E. BournePredicting the Polypharmacology of Drugs: Identifying New Uses through Chemoinformatics, Structural Informatics, and Molecular Modeling-Based Approaches 163-205. [CrossRef]
6. G. Panagiotou, O. Taboureau. 2012. The impact of network biology in pharmacology and toxicology. *SAR and QSAR in Environmental Research* **23**:3-4, 221-235. [CrossRef]
7. YUN LI, KANG TU, SIYUAN ZHENG, JINGFANG WANG, YIXUE LI, PEI HAO, XUAN LI. 2011. ASSOCIATION OF FEATURE GENE EXPRESSION WITH STRUCTURAL FINGERPRINTS OF CHEMICAL COMPOUNDS. *Journal of Bioinformatics and Computational Biology* **09**:04, 503-519. [CrossRef]
8. Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, Roded Sharan. 2011. Combining Drug and Gene Similarity Measures for Drug-Target Elucidation. *Journal of Computational Biology* **18**:2, 133-145. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links] [Supplemental Material]
9. Philip E. Bourne, Lei Xie, Li Xie. 2011. Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. *Annual Review of Pharmacology and Toxicology* **52**:1, 110301101444027. [CrossRef]
10. Edward J. Perkins, J. Kevin Chipman, Stephen Edwards, Tanwir Habib, Francesco Falciani, Ronald Taylor, Graham Van Aggelen, Chris Vulpe, Philipp Antczak, Alexandre Loguinov. 2011. Reverse engineering adverse outcome pathways. *Environmental Toxicology and Chemistry* **30**:1, 22-38. [CrossRef]
11. Christian Genest, Johanna Nešlehová, Noomen Ben Ghorbal. 2010. Spearman's footrule and Gini's gamma: a review with complements. *Journal of Nonparametric Statistics* **22**:8, 937-954. [CrossRef]
12. F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, D. di Bernardo. 2010. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences* **107**:33, 14621-14626. [CrossRef]
13. T. Mazza, A. Romanel, F. Jordan. 2010. Estimating the divisibility of complex biological networks by sparseness indices. *Briefings in Bioinformatics* **11**:3, 364-374. [CrossRef]
14. S. I. Berger, R. Iyengar. 2009. Network analyses in systems pharmacology. *Bioinformatics* **25**:19, 2466-2472. [CrossRef]