JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 2, 2012

Pp. 163–174
DOI: 10.1089/cmb.2011.0264

Large-Scale Elucidation of Drug Response Pathways in Humans

YAEL SILBERBERG,[1] ASSAF GOTTLIEB,[2] MARTIN KUPIEC,[1]
EYTAN RUPPIN,[2,3] and RODED SHARAN [2]

ABSTRACT

Elucidating signaling pathways is a fundamental step in understanding cellular processes and developing new therapeutic strategies. Here we introduce a method for the large-scale elucidation of signaling pathways involved in cellular response to drugs. Combining drug targets, drug response expression profiles, and the human physical interaction network, we infer 99 human drug response pathways and study their properties. Based on the newly inferred pathways, we develop a pathway-based drug-drug similarity measure and compare it to two common, gold standard drug-drug similarity measures. Remarkably, our measure provides better correspondence to these gold standards than similarity measures that are based on associations between drugs and known pathways, or on drug-specific gene expression profiles. It further improves the prediction of drug side effects and indications, elucidating specific response pathways that may be associated with these drug properties. Supplementary Material for this article is available at www.liebertonline.com/cmb.

Key words: algorithms, gene networks, graphs and networks, paths.

1. INTRODUCTION

Uncovering drug-induced signaling pathways is a key step in elucidating the mode of action of drugs and inferring other drug properties such as indications or side effects. Public databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010), BioCarta (Nishimura, 2001), and Reactome (Matthews et al. 2009) hold a collection of manually curated signaling pathways; however, the elucidation of drug-induced signaling pathways is only at its infancy. A number of studies have utilized the known curated pathways to gain insights on drugs' modes of action. Huang et al. (2005) combined data on gene expression and sensitivity to drugs in order to relate biological pathways derived from KEGG and BioCarta to various drugs. Chen et al. (2010) proposed an algorithm for extracting drug response-enriched sub-pathways from KEGG pathways, exploiting drug-specific gene expression data from the Connectivity Map (CMap) (Lamb, 2007). The aforementioned studies are all limited to using known curated pathways. Other studies have utilized protein interaction networks in order to uncover novel pathways underlying cellular response to perturbations (Bromberg et al., 2008; Yeger-Lotem et al., 2009; Yosef et al., 2009).

[1]Department of Molecular Microbiology and Biotechnology, [2]The Blavatnik School of Computer Science, and [3]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel.

Here, we attempt to infer novel drug-associated signaling pathways on a genome-scale. By integrating human protein-protein and protein-DNA interactions, drug targets and drug-induced gene expression data, we constructed 428 drug-specific signaling subnetworks and derived from them 99 putative signaling pathways that are supported by multiple subnetworks of this collection. The large majority (88) of these pathways do not significantly intersect known pathways from the KEGG database. We validated those pathways using literature-curated drug-gene interactions, and used them to define a novel similarity measure among drugs. Our new similarity measure corresponds better to gold standard drug-drug similarities (based on the commonly used Anatomical, Therapeutic, Chemical [ATC] classification system and chemical similarity) than to similarity measures constructed from curated pathways or from gene expression alone. Finally, we utilized the constructed pathway-based similarity measure to predict drug indications and side effects, validating these predictions against independent sources. Complementary to these predictions, we identified pathways associated with specific drug indications and side effects, demonstrating their potential role in the mode of action of the corresponding drugs.

2. RESULTS

A method for inferring drug response pathways

We devised a new method for elucidating signaling pathways that underlie drug response in humans. Our approach consists of two phases: (i) construction of drug-specific subnetworks by integrating protein interactions with drug-response gene expression data; and (ii) a meta-analysis of the inferred drug-response subnetworks to elucidate probable signaling pathways that are supported by multiple subnetworks (Fig. 1).

In the first phase, we assembled a human physical interaction network, comprised of protein-protein interactions (PPIs) and protein-DNA interactions (PDIs) obtained from multiple sources. Each physical interaction was assigned a confidence score based on the technologies by which it was detected and the number of times it was reproduced (see Methods). This network served as a basis for the subnetwork reconstruction. The latter was based on the assumption that, upon administration of a drug, its targets initiate a cascade of signaling events along the network that ultimately generate a measurable gene expression response (Bromberg et al., 2008). Accordingly, we inferred for each drug the most likely subnetwork connecting its targets to its induced set of differentially expressed genes, computed from drug response gene expression profiles extracted from the Connectivity Map database (Lamb, 2007). The inference aimed at maximizing the likelihood of the resulting subnetwork while minimizing the lengths of its constituent pathways (see Methods); following the work of Yeang et al. (2004), each such pathway, connecting a drug target to a differentially expressed gene, was required to end with a protein-DNA interaction (which mediates the effect of the drug on the expression of the target gene). Overall, we obtained 428 drug-specific subnetworks with an average size of 19 proteins per subnetwork.

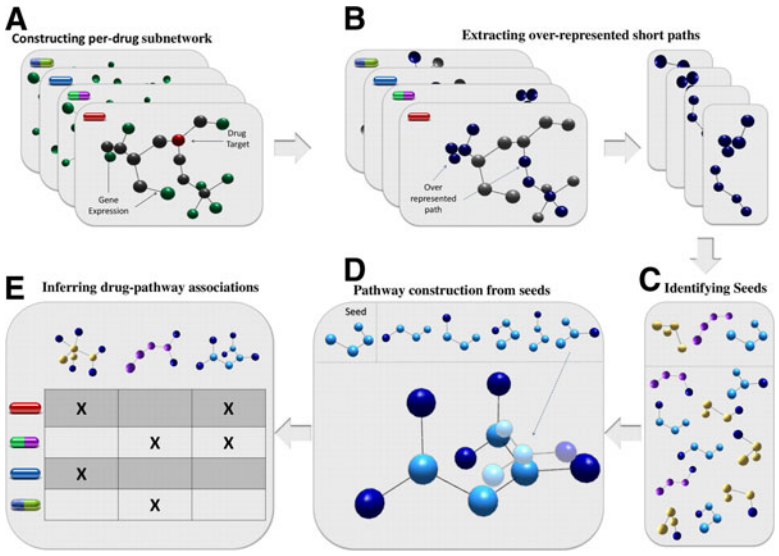

FIG. 1. Algorithmic pipeline. The pipeline includes five steps: **(A)** Constructing drug-specific subnetworks by compactly connecting drug targets to corresponding sets of differently expressed genes. **(B)** Identifying over-represented short paths of four proteins. **(C)** Constructing a non-redundant set of seed paths. **(D)** Seed expansion by overlapping paths to yield the full pathways. **(E)** Associating drugs with pathways if the pathway's proteins significantly overlap the drug-specific subnetwork.

For the second, meta-analysis phase, we devised a three-step agglomerative process in which we (i) identified prevalent short paths of length three (i.e., consisting of four proteins) in the drug-specific subnetworks; (ii) filtered the resulting set of paths for overlaps, obtaining a set of pathway seeds; and (iii) merged overlapping paths with seeds to obtain the final pathways. Full details are provided in the Methods section. The process resulted in a collection of 99 inferred pathways with an average size of 12 proteins per pathway (Table S1) (Supplementary Material is available at www.liebertonline.com/cmb).

Properties of the inferred pathways

The size and composition of the inferred pathways span a wide range (Fig. 2). Altogether, the pathways include 323 proteins, 65% of which appear in more than one pathway. We next applied the DAVID tool (Huang et al., 2009) to compute enrichment of GO terms and KEGG pathways (FDR correction is computed by DAVID by accounting for the entire set of GO terms or KEGG pathways, respectively). The most highly enriched GO terms belong to ''*response to stimuli*'' (57 enrichments; p-value $< 2e^{-16}$; e.g., ''inflammatory response'' and ''response to organic substance'') and ''*regulation mechanism*'' (229 enrichments, p-value = 0; e.g., ''regulation of cell death'' and ''positive regulation of RNA metabolic process''), as one would intuitively expect. Table S2 lists all enriched GO terms (False Discovery Rate [FDR] < 0.01) (Supplementary Material is available at www.liebertonline.com/cmb). Eighty-eight of our inferred pathways do not significantly overlap any known KEGG pathways (FDR < 0.01) and, henceforth, are termed *putative novel pathways* (PNPs). To assess the PNPs against current knowledge, we downloaded a set of manually curated pathways from the KEGG database (Kanehisa et al., 2010), retaining pathways with at least five proteins appearing in our physical network. Focusing on signaling pathways, we further filtered metabolic pathways from the analysis, resulting in 142 non-metabolic KEGG pathways. Analyzing the 299 proteins spanning the PNPs, we found that their frequency across inferred pathways is rank-correlated with their frequency across KEGG pathways (Spearman rank correlation, $r = 0.7$). The proteins spanning known pathways cover 72% of the PNPs, whereas only 39% ($\pm 3\%$) of a randomly selected set of proteins of the same size ($p < 0.01$). Figure 3 displays pathway similarities between PNPs and known pathways and among known pathways, based on shared proteins or interactions. The most abundant proteins in PNPs are MYC and TP53, appearing in more than 25% of the pathways. Expectedly, these proteins are also abundant in known pathways, appearing in the top 3% percentile and covering 10% of KEGG's non metabolic pathways.

Validation of the inferred pathways

In order to validate the inferred pathways, we first computed associations between drugs and pathways. We associated an inferred pathway with a drug if the corresponding drug-specific subnetwork (inferred from the gene expression response to the drug) was enriched with the pathway's proteins (see Methods). This yielded a pathway association profile characterizing each drug. Figure S1 displays a histogram of the number of drugs associated with each pathway (Supplementary Material is available at www.liebertonline.com/cmb). As a specific case study, we found that the set of proteins spanning the inferred pathways associated with the Dexamethasone drug was enriched in a list of sub-pathways extracted from KEGG by Chen et al. (2010) for the same drug ($p < 3e^{-3}$) (see Methods).

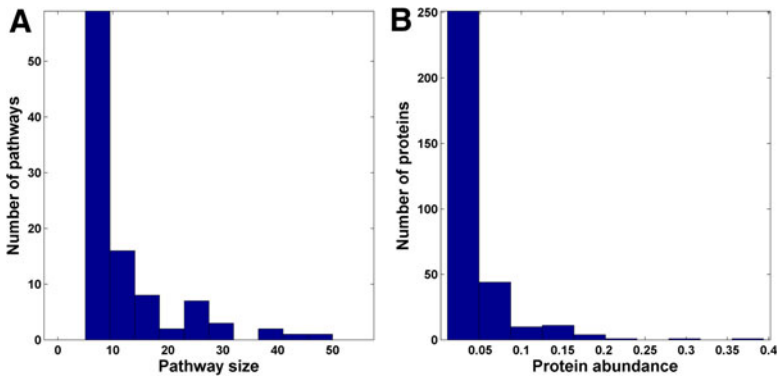

FIG. 2. Pathway properties. **(A)** Distribution of inferred pathways sizes. **(B)** Distribution of protein abundance in inferred pathways.

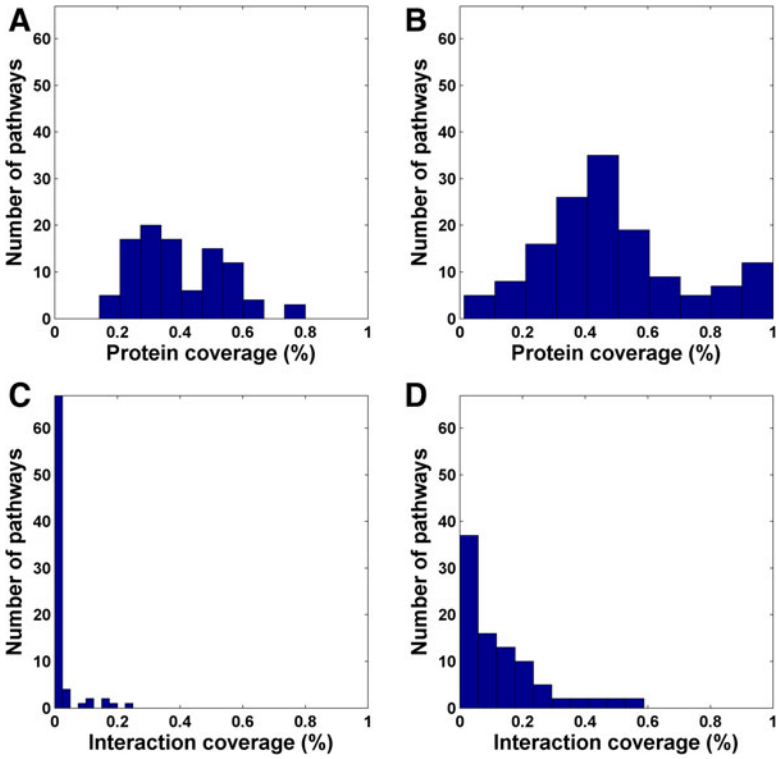

FIG. 3. Pathway similarity based on joint proteins between inferred and known pathways **(A),** and among known pathways **(B).** Pathway similarity based on joint interactions between inferred and known pathways **(C),** and among known pathways **(D).**

To assess the quality of pathways, we exploited a set of literature-curated drug-protein associations from PharmGKB (Klein et al., 2001). We built a gold standard set of drug-pathway associations, associating a drug with a pathway if the set of proteins belonging to the pathway significantly overlapped the literature-curated list of proteins associated with the drug (hypergeometric enrichment, FDR < 0.01). A total of 148 of the inferred pathway associations were supported by this type of analysis as drug-related ($p < 0.01$, in comparison to random protein sets of the same size; see Methods).

Next, we derived a pathway-based similarity assessment method between drugs using the Jaccard score (Jaccard, 1908). That is, the similarity between two drugs is the size of the intersection between their associated pathway sets over the size of the union of those sets. We compared our pathway-based drug-drug similarity (PDDS) measure to two types of well-established drug-drug similarity measures: (i) chemical similarity; and (ii) similarity based on the ATC classification system (Skrbo et al., 2004). For evaluation purposes, we tested additional drug-drug similarity measures based on (i) PDDS, where drug-pathway associations are based on enrichment of inferred pathways in the raw CMap drug-specific gene expression profiles; (ii) the raw CMap drug-response gene expression profiles; (iii) enrichment of known, human-curated pathways in drug-specific gene expression profiles; and (iv) enrichment of known pathways in drug-specific subnetworks (see Methods). In order to remove possible biases stemming from inclusion of drug targets in our inferred pathways, we also performed these comparisons in a second set of drug-pathway associations, computed after removing drug targets from their corresponding subnetworks. The area (AUC) under the receiver-operating characteristic (ROC) curve obtained using PDDS, 0.81 and 0.69 versus chemical and ATC similarity, respectively, was superior to all other tested drug-drug similarities, including similarity measures based on selected classes of related KEGG pathways (signal transduction and diseases pathway classes; Table 1). Figure 4 displays the ROC curves of a selected set of drug-drug similarities, while Figure S2 displays the full comparison (Supplementary Material is available at www.liebertonline.com/cmb).

Utilizing the inferred pathways for predicting drug side effects and indications

We utilized the associations between drugs and inferred pathways for the prediction of drug side effects and indications. We extracted side effects for 260 drugs (out of 428) from the SIDER database (Kuhn et al.,

TABLE 1. COMPARISON OF DRUG SIMILARITIES BASED ON INFERRED AND KNOWN PATHWAYS

Pathway type	*Drug-pathway association method*	*Chemical similarity (AUC)*	*ATC similarity (AUC)*
Inferred pathways	Subnetwork enrichment	0.81	0.69
Inferred pathways (excluding drug targets)	Subnetwork enrichment	0.81	0.68
Inferred pathways	Gene expression enrichment	0.64	0.58
All KEGG pathways	Subnetwork enrichment	0.64	0.58
KEGG disease pathways	Subnetwork enrichment	0.66	0.57
KEGG signal transduction pathways	Subnetwork enrichment	0.63	0.57
KEGG pathways, excluding metabolic pathways	Subnetwork enrichment	0.66	0.59
KEGG pathways, excluding metabolic pathways	Gene expression enrichment	0.62	0.53
Gene expression based similarity	N/A	0.76	0.68

2010). Focusing on more prevalent side effects, we considered only side effects caused by at least five drugs, resulting in 557 side effects and 18,584 drug-side effect associations (Kuhn et al., 2010) (see Methods). To predict side effects, we trained support vector machine (SVM) classifiers on each side effect separately. The success of such predictors critically depends on the quality of the drug-drug similarity measures used, assuming that similar drugs tend to have similar side effects (Atias and Sharan, 2011) and, hence, may serve as an operative yardstick for the quality of the drug-pathway associations we have inferred. We assessed the resulting prediction accuracy using a 10-fold cross-validation scheme, obtaining a highly significant AUC score distribution (Wilcoxon rank sum test, $p < 3e^{-68}$; Fig. 5). Remarkably, the AUC distribution obtained using our drug-drug similarity measure has higher AUC mean and is better separated from the random model than that obtained using a drug-drug similarity measure relying on known pathways ($p < 4e^{-7}$; Fig. S3) (Supplementary Material is available at www.liebertonline.com/cmb). We further compared our side effect prediction scheme to that proposed by Atias and Sharan (2011). Out of 416 side effects predicted by both methods, 195 side effect predictions (47%) obtained higher AUC scores using our methodology (Fig. S3) (Supplementary Material is available at www.liebertonline.com/cmb). Using the inferred pathways, we found 242 side effects whose AUC score was significantly higher than those predicted by the random model (Wilcoxon ranked sum test, FDR < 0.01), compared to 229 such side effects when using KEGG pathways (see Methods). Next, focusing on these 242 side effects, we predicted their associations with 168 drugs with no prior side effect information in SIDER, inferring 7435 associations overall (Table S3) (Supplementary Material is available at www.liebertonline.com/cmb). We validated these predictions against textual description of drug health effects from the Hazardous Substances Data Bank (HSDB) database (Wexler, 2001) by applying a simple textual search scheme to query the association of these 242 side effects with each drug. We managed to retrieve 36% of the recorded drug-side

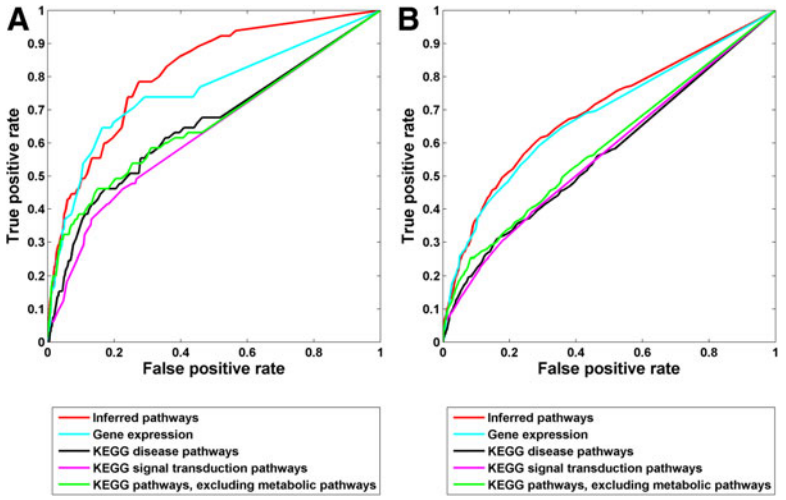

FIG. 4. ROC curves of drug-drug similarity measures. The ROC curves are computed using inferred pathways, known pathways and gene expression relative to chemical drug-drug similarity **(A)** and drug-drug similarity based on similar fourth level ATC classification **(B).**

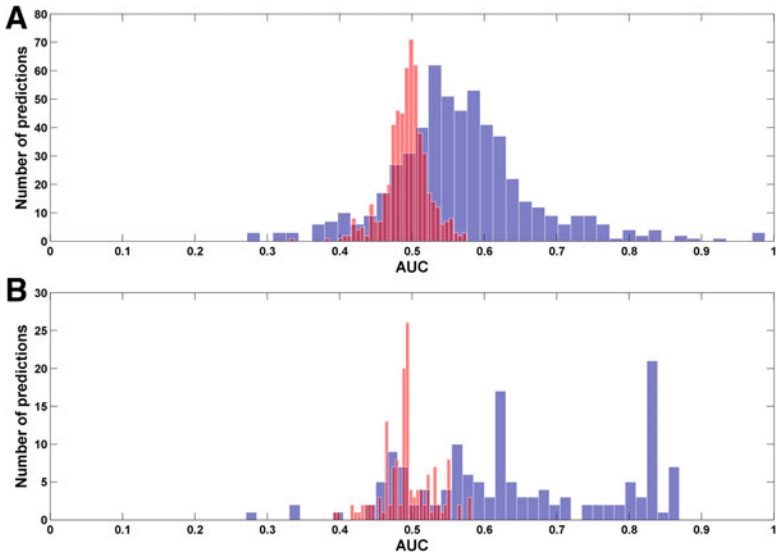

FIG. 5. Distribution of obtained AUCs in the prediction of side effects **(A)** and indications **(B),** while using drug-specific profiles of pathways (blue) relative to a random model (red).

effect associations across 80 drugs registered in HSDB, with an accuracy of 11% (hypergeometric test, $p < 2e^{-44}$).

We highlight several of these predictions: Zuclopenthixol, an antipsychotic drug, is predicted to cause priapism, a persistent, usually painful erection. Several reports indeed observed this side effect in patients taking this drug (Fishbain, 1985; Kihl et al., 1980; Salado et al., 2002; Van Hemert et al., 1995). Tenoxicam and Nimesulide, non-steroidal anti-inflammatory drugs (NSAIDs) of the oxicam type, are predicted to cause Stevens-Johnson syndrome, a potentially deadly skin disease that usually results from adverse drug reaction. Tenoxicam and Nimesulide were indeed reported to (rarely) produce such a side effect, as are many other NSAIDs (Mockenhaupt et al., 2007; Roujeau et al., 1995). Nimesulide is also predicted to cause a similar side effect of toxic epidermal necrolysis (TEN), as has been reported by Chatterjee et al. (2008).

To study the relations between pathways and side effects, we searched for side effects whose associated drugs are (hypergeometrically) enriched in the drug sets associated with each pathway. We found 82 enrichments of 42 side effects within 30 inferred pathways (FDR < 0.01; see Methods). Table S4 lists all the associations between pathways and side effects (Supplementary Material is available at www.liebertonline.com/cmb). For example, the AV block side effect is enriched in the pathway displayed in Figure 6A. This pathway contains the drug target SCN5A, a voltage-gated sodium channel subunit, primarily found in cardiac muscles. Mutations in SCN5A have been associated with multiple cardiac disorders including functional two-to-one AV block (Lupoglazoff et al., 2001). Additionally, ATF3, a transcription factor of the CREB family, is regulated in this pathway. Over-expression of ATF3 in the heart was reported to induce myocardial rearrangements, fibrosis and conduction abnormalities (van Veen et al., 2005); in addition, this gene was reported to be up-regulated in SCN5A transgenic mice (Royer et al., 2005). Encouragingly, we found that one of the drugs we predicted to cause an AV block, Tocainide, targets SCN5A. SCN5A interacts with syntrophin alpha 1, SNTA1, also interacting with NOS1. Mutation in

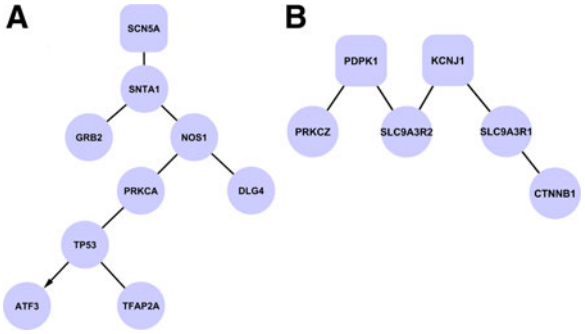

FIG. 6. Examples of pathways associated with side effects. Shown are inferred pathways that are enriched for AV Block **(A)** and microvascular complication of diabetes **(B).** Rectangles denote known drug targets; undirected edges denote protein-protein interactions, and directed edges denote protein-DNA interactions.

SNTA1 was associated with long QT syndrome, a disease involving abnormal repolarization of the heart which often coupled with abnormalities of atrioventricular conduction (Saoudi et al., 1991). SNTA1 is believed to provoke long QT syndrome by forming a SNTA1-based NOS1 complex which acts as a regulator of the cardiac sodium channel SCN5A (Ueda et al., 2008). A discussion of the two additional pathways that display side effect enrichment is provided in the Supplementary Material (Supplementary Material is available at www.liebertonline.com/cmb).

Next, we aimed to use the pathway-based information to predict drug indications. Gold standard drug indications were taken from Gottlieb et al. (2011), who assembled a comprehensive set of drug indications from different sources, including FDA labels, DrugBank indications (Wishart et al., 2008), and the drugs.com website (see Methods). We retained all indications that associate with at least five drugs in our data set, resulting in 1,669 associations between 283 drugs and 145 indications. We trained an SVM classifier, similar to the predictor of side effects described above, for those 145 indications and assessed the resulting prediction accuracy using a 10-fold cross-validation scheme. The distribution of the AUC scores obtained in cross-validation is significantly better than random (Wilcoxon rank sum test, $p < 7e^{-23}$; Fig. 5). We found 86 indications whose AUC scores were significantly higher than those predicted by a random model (FDR < 0.01), compared to 99 such indications when using KEGG pathways (see Methods). We considered only the 86 indications with significantly higher than random AUC score to predict 783 associations with 145 drugs that have no known indication in our data set (Table S5) (Supplementary Material is available at www.liebertonline.com/cmb). In order to validate these predictions, we checked for their appearance in currently available clinical trials. We downloaded drug indications currently studied in clinical trials (phases I–IV) from the registry of federally and privately supported clinical trials conducted around the world (http://clinicaltrials.gov) and employed the method described in (Gottlieb et al., 2011) to extract the drug-disease associations. We found 89 investigated drug indications involving one of the 145 drugs and 86 diseases included in our predictions. Our predictions cover 17% of these investigated drug indications ($p < 4e^{-4}$).

We further searched for enrichment of drug indications in specific pathways. We found 58 indications enriched within 16 inferred pathways (FDR < 0.01; see Methods and Table S6) (Supplementary Material is available at www.liebertonline.com/cmb). We highlight one of these pathways, displayed in Figure 6B. This pathway is enriched with microvascular complications of diabetes, including diabetic nephropathy, neuropathy, and retinopathy, the last being the most common complication (Fowler, 2008). The pathway includes two known drug targets: (i) 3-phosphoinositide dependent protein kinase-1 (PDPK1), and (ii) ATP-dependent potassium channel (KCNJ1). KCNJ1 is connected to β-catenin-1 (CTNNB1) through solute carrier family 9 (sodium/hydrogen exchanger), member 3 regulator 1 (SLC9A3R1). KCNJ1, mostly located in the kidney and pancreatic islets, is the principal target of Sulfonylurea, a class of drugs widely used to promote insulin secretion in the treatment of diabetes mellitus. Retinal levels and nuclear translocation of CTNNB1, a key effector in the canonical Wnt pathway, were increased in humans with diabetic retinopathy and in three diabetic retinopathy models (Chen et al., 2009). Finally, KCNJ1 is also connected via another solute carrier family 9 (SLC9A3R2) to the protein PDPK1, a master kinase, which is connected to its activated protein kinase C-zeta (PRKCZ) (Hodgkinson and Sale, 2002). It was suggested that the activation of PKC has an important role in the development of microvascular endothelial dysfunction during the early stages of diabetes (Yuan et al., 2000). A discussion of three additional pathways, enriched with asthma is provided in Supplementary Material (Supplementary Material is available at www.liebertonline.com/cmb).

3. DISCUSSION

We presented a new method for elucidating drug response pathways. Our method utilizes a physical network of PPIs and PDIs, drug target information, and drug response gene expression profiles to construct drug-specific subnetworks and subsequently infer signaling pathways supported by these subnetworks. By associating drugs with the reconstructed pathways, we created drug-pathway association profiles, which were validated against a literature-curated data source of drug-gene interactions. We further validated that drugs sharing chemical similarity or sharing the fourth level in ATC classification tend to be associated with similar pathways.

We utilized our inferred pathways to predict drug side effects and indications, validating the results in a cross-validation setting and against unseen data. Quite surprisingly perhaps, the inferred pathways lead to better predictors of side effects than those obtained by relying on the KEGG human-curated ensemble of

signaling pathways, demonstrating their potential power in predicting drug properties. Furthermore, we could tie the drug properties with specific pathways that significantly associate with them.

We note that the signaling pathways identified in this analysis were limited to those mediated by physical protein interactions. Incorporating other types of cell signaling, such as metabolic pathways, may be a subject for further work. Furthermore, our method is generic in the sense that other classes of stimuli, such as endogenous hormones, could be handled too, provided that information on the proteins that initiate the cellular response or are affected by it are available.

4. METHODS

Data sets

To reconstruct drug-specific subnetworks, we compiled physical network data, drug targets and drug response information. Human PPI network, spanning 10,393 proteins and 44,794 protein-protein interactions was compiled from (Breitkreutz et al., 2008; Ewing et al., 2007; Jiang et al., 2007; Rual et al., 2005; Stelzl et al., 2005; Xenarios et al., 2002). As the interaction measurements techniques are noisy, each interaction is assigned a confidence score that accounts for the experimental techniques by which the interaction was detected and the number of times the interaction was reproduced using a logistic regression model adapted from (Sharan et al., 2005). These confidence scores are used by our algorithm to determine the highest confidence pathways connecting the drug targets and the expressed genes, as described in Yosef et al. (2009). Human PDI was downloaded from TRED database (Jiang et al., 2007). TRED incorporates experimental data followed by exhaustive literature curation. As stated by the authors, hand curation was applied as a crucial part of the data collection to ensure data accuracy. Having only one source of PDIs, all interactions were assigned the maximal confidence score (1).

Drug targets were extracted from DrugBank (Wishart et al., 2008), DCDB (Liu et al., 2010), and KEGG DRUG (Kanehisa and Goto, 2000) databases. Drug-response gene expression data were downloaded from the Connectivity Map (CMap, build 02 [Lamb, 2007]). CMap contains 6,100 gene expression measurements in response to the administration of 1,309 drugs and small molecules. These measurements were taken under different drug concentrations and on different cell-line types using the Affymetrix HG-U133A and HT-HG-U133A Array. In order to form drug-specific signatures, we followed the normalization and filtering procedures described in Iskar et al. (2010), resulting in gene expression profiles of 1,144 drugs in three human cell lines (human promyelocytic leukemia cell line [HL60], human breast adenocarcinoma cell line [MCF7], and human prostate cancer cell line [PC3]).

For validation and application purposes, we collected manually curated information on drugs, including drug-gene interactions, chemical structure, categorization, side effects, and indications. Drug-gene associations were downloaded from PharmGKB (Klein et al., 2001). Canonical simplified molecular input line entry specification (SMILES) (Weininger, 1988) of drugs and the ATC codes (Skrbo et al., 2004) were extracted from DrugBank (Wishart et al., 2008). Drug side effects were downloaded from the SIDER database (Kuhn et al., 2010). Drug indications were computed according to the method presented in Gottlieb et al. (2011). In brief, they assembled associations between diseases listed in the Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2002) and their indicated drugs, registered in DrugBank (Wishart et al., 2008). Initially, OMIM disease names were mapped to Unified Medical Language System (UMLS) concepts (Bodenreider, 2004) using a natural language processing tool named MetaMap (Aronson, 2001). Subsequently, drug-disease associations were extracted from the following sources: (i) links between concepts and drugs embedded in the UMLS; (ii) drug indications appearing in the http://drugs.com website; and (iii) UMLS concepts extracted from FDA package inserts and indications registered in DrugBank.

Constructing drug-specific subnetworks

For each drug, we inferred the most likely subnetwork that connects the drug's targets (source proteins) to the set of genes whose expression significantly changed in response to the drug (target proteins). In order to build a drug-specific gene expression signature (i.e., target list), we computed such a signature for every drug in every cell line by computing a z-score for each gene relative to the mean and standard deviation of all the genes in the drug's microarray, choosing significantly differentially expressed genes (z-score corrected for FDR 0.01). Finally, a drug-specific expression profile is the union of expression profiles extracted from each of the three cell lines.

Drug-specific subnetworks were reconstructed using the method of Yosef et al. (2009), which provides a trade-off between constructing a globally optimal subnetwork, connecting the sources and the targets in the most compact manner, and constructing a locally optimal subnetwork connecting the sources and targets via shortest paths. Since the targets were inferred from gene expression data, we further constrained the last interaction in each source-to-target route to be a protein-DNA interaction, connecting a protein (transcription factor) to a transcribed gene, decoupled from its encoded protein (i.e., genes and their corresponding encoded proteins are assigned different nodes in the network, where genes are connected to proteins solely through PDIs). This reconstruction strategy yielded 428 drug-specific signaling subnetworks.

Inferring pathways from drug-specific subnetworks

The subsequent pathway inference was performed in four consecutive steps. First, we enumerated all short paths of length three appearing in any of the drug-specific subnetworks and computed their frequency across the subnetworks. In order to obtain pathways representing a typical drug response, we filtered short paths that appear in a single drug-specific subnetwork. We verified that filtering short paths appearing in two drug-specific subnetworks as well as removing this filtering step altogether, resulted in minor change in performance (AUC diff <0.02 for validation against chemical similarity and AUC $<$ 0.01 for validation against ATC similarity). We constructed 20 random subnetworks per drug by connecting the same sources (drug targets) with an equal-sized set of randomly chosen genes as targets. We further filtered for short paths appearing in more than 5% of the randomly constructed subnetworks. We validated that stricter filtering, removing paths appearing in over 1% of random subnetwork resulted in minor changes (AUC diff <0.01 for validation against chemical similarity and AUC $< e^{-4}$ against ATC similarity). The final set consisted of 3,416 prevalent short pathways, sorted in decreasing order of their frequency.

In the second step, we iteratively started from the most frequent short pathway (termed a *seed*) and filtered all short pathways sharing at least one PPI or PDI edge with the seed. This iterative process produced 128 seeds. The third step expands the seeds to form pathways by joining the short pathways filtered in the previous step to the growing seeds if they share at least two edges. The fourth and final step merges smaller pathways into larger ones if they share over 75% of their proteins, the merging process is repeated until converge; unexpanded seeds are removed. The four-step process yielded a collection of 99 inferred pathways with an average size of 12 proteins each.

Validating the inferred pathways

A drug is associated with an inferred pathway if the proteins spanning the drug-specific subnetwork significantly overlap the set of proteins spanning the inferred pathway (hypergeometric enrichment, FDR < 0.01), creating a pathway association profile for each drug. For comparison purposes, we also computed drug-pathway associations that are based on the expression profiles only or associations between drugs and known KEGG pathways. In order to compute the enrichment of pathway proteins in a drug-specific gene expression profile, we computed a *t*-test (Student, 1908a) between the gene expression values of the pathway genes and all other genes in the expression profile. A drug-pathway association was assigned the minimal *t*-test *p*-value obtained in each of the three cell lines.

To validate the pathways, we exploited a set of literature-curated drug-protein interactions from PharmGKB (Klein et al., 2001). We built a gold standard set of drug-pathway associations, by associating a drug with a pathway if the set of proteins included in that pathway significantly overlapped the curated list of proteins associated with the drug, removing drug targets from the PharmGKB drug-gene interaction to avoid potential bias (hypergeometric enrichment, FDR < 0.01). To assess the number of drug-pathway associations expected at random, we permuted the drug-gene associations 100 times, while maintaining the number of genes associating with each drug.

To further validate the inferred pathways, we constructed a pathway-based drug-drug similarity measure (PDDS), computed using the Jaccard score (Jaccard, 1908) between pathway profiles of drug pairs. We compared our PDDS to two gold standard drug-drug similarity measures: (i) chemical similarity and (ii) identity of fourth level ATC classification, identifying the pharmacological, therapeutic and chemical subgroup of the drug. To measure the chemical similarity between drugs, we computed hashed fingerprints for each drug using the Chemical Development Kit (CDK) with default parameters (Steinbeck et al., 2006)

on the drug Canonical SMILES (Weininger, 1988). The similarity score between a pair of drugs is computed on their fingerprints according to the two-dimensional Tanimoto coefficient (Tc) (Tanimoto, 1957). Following (Matter, 1997), we considered two drugs to be chemically similar if their Tc was above 0.85. Using lower Tc thresholds achieved similar AUCs (AUC difference <0.004 for Tc as low as 0.7). In order to validate that our results are not biased due to known chemical similarities between drugs sharing similar targets, we verified that removing drug targets from the drug-specific subnetworks before computation of the PDDS resulted in a similar AUC (AUC difference <0.01).

Prediction of side effects

We predicted side effects for 174 drugs that are not represented in SIDER by training a polynomial SVM classifier (degree of three) for each side effect using the log *p*-values of the drug-pathway associations as features. We assessed our prediction accuracy by employing a 10-fold cross validation scheme for each side effect (averaged over 10 independent runs) and compared the obtained AUCs to those formed on 10 randomized data sets where the drug side-effect associations were permuted, while maintaining the number of drug associations for each side effect. We tested the performance obtained when using different degrees (degree ranging between 2 and 5), a different cost parameter (cost parameter of 0.01, 1, and 100) as well as using a linear SVM, all of which did not significantly alter the results (mean AUC diff <0.02). A similar prediction scheme was employed using associations between drugs and known pathways. To study the relations between pathways and side effects, we computed hypergeometric enrichment of side effects in pathways. In order to correct for multiple hypothesis testing, we used an empirical scheme: we permuted 100 times the drug-pathway associations (maintaining side effect degrees), and recorded for each side effect its minimal enrichment *p*-value under each of the permutations. We set the significance cutoff to the bottom 1% of the recorded *p*-values.

ACKNOWLEDGMENTS

Y.S. and A.G. were partially funded by the Edmond J. Safra Bioinformatics Program. E.R. and R.S. were partially supported by a Bikura grant from the Israel Science Foundation. R.S. was further supported by a research grant from the Israel Science Foundation (grant 241/11). M.K. was supported by grants from the Israeli Ministry of Science and Technology, by the Israel Cancer Research Fund, and by the Israel Cancer Association.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aronson, A.R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA* Symp. 17–21.

Atias, N., and Sharan, R. 2011. An algorithmic framework for predicting side effects of drugs. *J. Comput. Biol.* 18, 207–218.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 32, D267–D270.

Breitkreutz, B.J., Stark, C., Reguly, T., et al. 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*. 36, D637–D640.

Bromberg, K.D., Ma'ayan, A., Neves, S.R., et al. 2008. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science* 320, 903–909.

Chatterjee, S., Pal, J., and Biswas, N. 2008. Nimesulide-induced hepatitis and toxic epidermal necrolysis. *J. Postgrad Med.* 54, 150–151.

Chen, X., Xu, J., Huang, B., et al. 2010. A sub-pathway-based approach for identifying drug response principal network. *Bioinformatics* 27, 649–654.

Chen, Y., Hu, Y., Zhou, T., et al. 2009. Activation of the Wnt pathway plays a pathogenic role in diabetic retinopathy in humans and animal models. *Am. J. Pathol.* 175, 2676–2685.

Ewing, R.M., Chu, P., Elisma, F., et al. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3, 89.

Fishbain, D.A. 1985. Priapism resulting from fluphenazine hydrochloride treatment reversed by diphenhydramine. *Ann. Emerg. Med.* 14, 600–602.

Fowler, M.J. 2008. Microvascular and macrovascular complications of diabetes. *Clin. Diabetes* 26, 77.

Gottlieb, A., Stein, G.Y., Ruppin, E., et al. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496.

Hamosh, A., Scott, A.F., Amberger, J., et al. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52–55.

Hodgkinson, C.P., and Sale, G.J. 2002. Regulation of both PDK1 and the phosphorylation of PKC-zeta and -delta by a C-terminal PRK2 fragment. *Biochemistry* 41, 561–569.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1.

Huang, R., Wallqvist, A., Thanki, N., et al. 2005. Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J.* 5, 381–399.

Iskar, M., Campillos, M., Kuhn, M., et al. 2010. Drug-induced regulation of target expression. *PLoS Comput. Biol.* 6, e1000925.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. *Bul. Soc. Vaudoise Sci. Nat.* 44, 223–270.

Jiang, C., Xuan, Z., Zhao, F., et al. 2007. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.* 35, D137–D140.

Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30.

Kanehisa, M., Goto, S., Furumichi, M., et al. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360.

Kihl, B., Bratt, C.G., Knutsson, U., et al. 1980. Priapsim: evaluation of treatment with special reference to saphenocavernous shunting in 26 patients. *Scand. J. Urol. Nephrol.* 14, 1–5.

Klein, T.E., Chang, J.T., Cho, M.K., et al. 2001. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* 1, 167–170.

Kuhn, M., Campillos, M., Letunic, I., et al. 2010. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6, 343.

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

Liu, Y., Hu, B., Fu, C., et al. 2010. DCDB: Drug Combination Database. *Bioinformatics* 26, 587–588.

Matter, H. 1997. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* 40, 1219–1229.

Matthews, L., Gopinath, G., Gillespie, M., et al. 2009. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37, D619–D622.

Mockenhaupt, M., Viboud, C., Dunant, A., et al. 2007. Stevens–Johnson syndrome and toxic epidermal necrolysis: assessment of medication risks with emphasis on recently marketed drugs. The EuroSCAR-study. *J. Invest. Dermatol.* 128, 35–44.

Nishimura, D. 2001. A view from the web: BioCarta. *Biotech Software Internet Rep.* 2, 117–120.

Roujeau, J.C., Kelly, J.P., Naldi, L., et al. 1995. Medication use and the risk of Stevens-Johnson syndrome or toxic epidermal necrolysis. *N. Engl. J. Med.* 333, 1600–1607.

Rual, J.F., Venkatesan, K., Hao, T., et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.

Salado, J., Blazquez, A., Diaz-Simon, R., et al. 2002. Priapism associated with zuclopenthixol. *Ann. Pharmacother.* 36, 1016–1018.

Saoudi, N., Bozio, A., Kirkorian, G., et al. 1991. Prolonged QT, atrioventricular block, and sudden death in the newborn: an electrophysiologic evaluation. *Eur. Heart. J.* 12, 838–841.

Sharan, R., Suthram, S., Kelley, R.M., et al. 2005. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* 102, 1974.

Skrbo, A., Begovic, B., and Skrbo, S. 2004. [Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes]. *Med. Arh.* 58, 138–141.

Steinbeck, C., Hoppe, C., Kuhn, S., et al. 2006. Recent developments of the chemistry development kit (CDK)—an open-source Java library for chemo- and bioinformatics. *Curr. Pharm. Des.* 12, 2111–2120.

Stelzl, U., Worm, U., Lalowski, M., et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.

Student, Z. 1908a. The probable error of a mean. *Biometrica* 6, 1–25.

Tanimoto, T.T. 1957. *IBM Internal Report* [November 17, 1957].

Ueda, K., Valdivia, C., Medeiros-Domingo, A., et al. 2008. Syntrophin mutation associated with long QT syndrome through activation of the nNOS–SCN5A macromolecular complex. *Proc. Natl. Acad. Sci. USA* 105, 9355.

Van Hemert, A.M., Meinhardt, W., Moehadjir, D., et al. 1995. Recurrent priapism as a side effect of zuclopenthixol decanoate. *Int. Clin. Psychopharmacol.* 10, 199.

van Veen, T.A., Stein, M., Royer, A., et al. 2005. Impaired impulse propagation in Scn5a-knockout mice: combined contribution of excitability, connexin expression, and tissue architecture in relation to aging. *Circulation* 112, 1927–1935.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36.

Wexler, P. 2001. TOXNET: an evolving web resource for toxicology and environmental health information. *Toxicology* 157, 3–10.

Wishart, D.S., Knox, C., Guo, A.C., et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. 36, D901–906.

Xenarios, I., Salwinski, L., Duan, X.J., et al. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305.

Yeang, C.H., Ideker, T., and Jaakkola, T. 2004. Physical network models. *J. Comput. Biol.* 11, 243–262.

Yeger-Lotem, E., Riva, L., Su, L.J., et al. 2009. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* 41, 316–323.

Yosef, N., Ungar, L., Zalckvar, E., et al. 2009. Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.* 5, 248.

Yuan, S.Y., Ustinova, E.E., Wu, M.H., et al. 2000. Protein kinase C activation contributes to microvascular barrier dysfunction in the heart at early stages of diabetes. *Circ. Res.* 87, 412–417.

Address correspondence to:
Dr. Roded Sharan
The Blavatnik School of Computer Science
Tel-Aviv University
Tel-Aviv, Israel 69978

E-mail: roded@post.tau.ac.il