Molecular Systems Biology 7; Article number 496; doi:10.1038/msb.2011.26
Citation: ***Molecular Systems Biology* 7:496**

www.molecularsystemsbiology.com

molecular systems biology

PREDICT: a method for inferring novel drug indications with application to personalized medicine

Assaf Gottlieb[1], Gideon Y Stein[2,3], Eytan Ruppin[1,2] and Roded Sharan[1,*]

[1] The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, [2] Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel and [3] Department of Internal Medicine 'B', Beilinson Hospital, Rabin Medical Center, Petah-Tikva, Israel
* Corresponding author. The Blavatnik School of Computer Science, Tel-Aviv University, Haim-Levanon, Tel-Aviv 69978, Israel.
Tel.: +972 3 640 7139; Fax: +972 3 640 9357; E-mail: roded@post.tau.ac.il

Received 12.1.11; accepted 12.4.11

Inferring potential drug indications, for either novel or approved drugs, is a key step in drug development. Previous computational methods in this domain have focused on either drug repositioning or matching drug and disease gene expression profiles. Here, we present a novel method for the large-scale prediction of drug indications (PREDICT) that can handle both approved drugs and novel molecules. Our method is based on the observation that similar drugs are indicated for similar diseases, and utilizes multiple drug–drug and disease–disease similarity measures for the prediction task. On cross-validation, it obtains high specificity and sensitivity (AUC=0.9) in predicting drug indications, surpassing existing methods. We validate our predictions by their overlap with drug indications that are currently under clinical trials, and by their agreement with tissue-specific expression information on the drug targets. We further show that disease-specific genetic signatures can be used to accurately predict drug indications for new diseases (AUC=0.92). This lays the computational foundation for future personalized drug treatments, where gene expression signatures from individual patients would replace the disease-specific signatures.
Molecular Systems Biology **7**:496; published online 7 June 2011; doi:10.1038/msb.2011.26
Subject Categories: bioinformatics; molecular biology of disease
Keywords: drug indication prediction; drug repositioning; drug repurposing; machine learning; personalized medicine

Introduction

Associating accurate indications with new molecules or alternative indications for approved drugs is a key step in drug development. High drug development costs (DiMasi *et al*, 2003) call for computational solutions that would minimize production time and ultimately development costs (Terstappen and Reggiani, 2001). Current computational methods for indication prediction mainly focus on small-scale applications, where drugs that target proteins in disease-specific molecular networks are sought (Kinnings *et al*, 2009; Li *et al*, 2009; Kotelnikova *et al*, 2010), while large-scale attempts are still scarce.

Previous attempts for large-scale identification of novel drug indications include: (i) matching of gene expression profiles proposed by the Connectivity Map (CMap) consortium (Lamb *et al*, 2006; see also Hu and Agarwal (2009)) and (ii) the 'Guilt by Association' (GBA) approach (Chiang and Butte, 2009). CMap is a database containing ranked drug response gene expression profiles. Querying the database with a disease-specific genetic signature, CMap identifies drug response profiles that either correlate (i.e., upregulated signature genes tend to appear at the top of the profile while down-regulated signature genes tend to appear at the bottom of the profile) or anti-correlate with it. A similar approach was proposed by Hu and Agarwal (2009), using gene expression measurements downloaded from the Gene Expression Omnibus (GEO; Edgar *et al*, 2002). While the CMap approach can be applied to any potential drug, its prediction power has not been assessed at large scale to date. As discussed in Hu and Agarwal (2009), the gene expression approach currently suffers both from low precision due to profiles generated under different conditions and from incapability to capture drug–disease associations that are not manifested at the gene expression level. GBA (Chiang and Butte, 2009) attempts to predict novel associations between drugs and diseases by assuming that if two diseases are treated by the same drug, alternative drugs treating only one of them might treat also the other. It is thus applicable only in the drug repositioning setting, where some indication for the drug in question is

already known. A related work combined drug–drug similarities and drug indications for the inference of drug–gene associations (Hansen *et al*, 2009).

Here, we present an approach for predicting novel associations between drugs and diseases that can operate on both drugs with approved indications and on novel molecules with no indication information. Given a query association, we measure the similarity of the pertaining drug and disease to drug–disease pairs that are known to be associated, and rank the accumulative evidence for association using a logistic regression scheme. The prediction process is aided by a comprehensive drug–disease association data set that we have compiled and a collection of novel drug–drug similarity measures. Importantly, we show the potential utility of our approach also in a personalized medicine setting, in which a disease name is replaced by a gene expression signature; and consequently, disease–disease similarity is measured via the similarity of the corresponding signatures.

Results and discussion

PREDICT—an algorithm for predicting drug indications

We designed a novel algorithm for PREdicting Drug IndiCations (PREDICT). Given a gold standard set of drug–disease associations (*known associations*), the algorithm ranks additional drug–disease associations based on their similarity to the known associations. The algorithm works in three phases (Figure 1): (i) construction of drug–drug and disease–disease similarity measures; (ii) exploiting these similarity measures to construct classification features and subsequent learning of a classification rule that distinguishes true from false drug–disease associations; and (iii) application of the classifier to predict new associations. The gold standard of drug–disease associations used for training was constructed from multiple sources, matching drugs, drug indications and disease names using the Unified Medical Language System (UMLS), as described in Materials and methods. In brief, the gold standard data set spans 1933 associations between 593 drugs taken from DrugBank (Wishart *et al*, 2008) and 313 diseases listed in the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al*, 2002; Supplementary Table S1). OMIM holds a comprehensive set of phenotypic descriptions of diseases and disorders, including complex multiple-gene disorders, allowing for the construction of phenotypic similarity measures. Hard to treat congenital anomalies were removed using the International Classification of Diseases (ICD-10) (see Materials and methods). We provide an overview of the algorithmic steps below; full details are given in Materials and methods.

For the first algorithmic phase, we assembled five drug–drug similarity measures between the 593 drugs in our gold standard and two disease–disease similarity measures between the 313 associated diseases. The drug–drug similarity measures include chemical similarity, similarities based on registered and predicted side effects (Kuhn *et al*, 2010; Atias and Sharan, 2011) and similarities between drug targets, including sequence similarity, distance on a protein–protein interaction (PPI) network and gene ontology (GO) (Ashburner

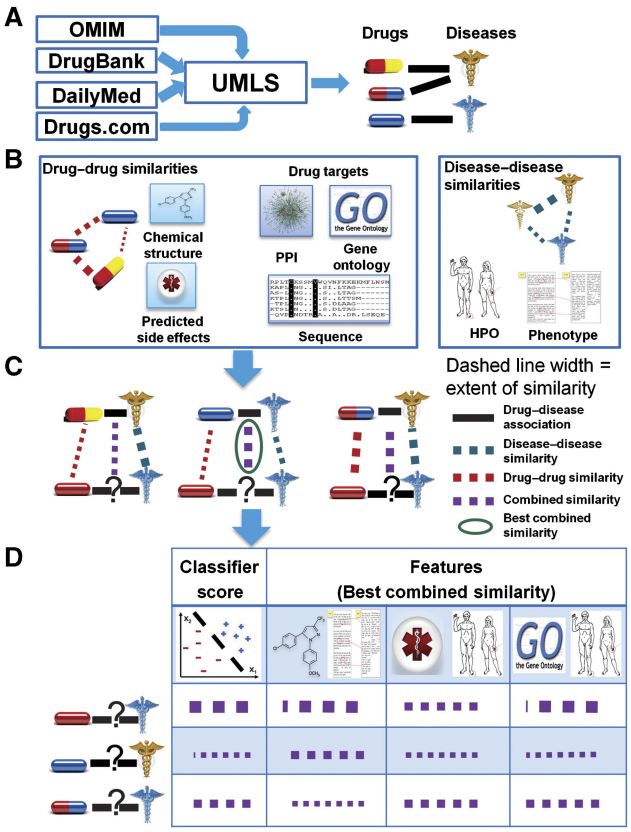

Figure 1 Algorithmic pipeline: formation of drug–disease associations (**A**), creation of drug–drug and disease–disease similarity metrics (**B**), scoring possible drug indications according to their similarity to known drug indications (**C**) and integration of the similarities to classification features and subsequent classification (**D**).

et al, 2000) semantic similarity. The disease–disease similarity measures are based on semantic similarity of disease phenotypes according to the van Driel *et al* (2006) text mining scheme and the human phenotype ontology (HPO; Robinson and Mundlos, 2010).

The second algorithmic phase integrates the drug–drug similarities and disease–disease similarities to construct classification features and subsequently learns a classification rule that distinguishes between true and false drug–disease associations. For each query drug–disease association, we constructed features expressing its similarity to the closest known drug–disease association, using the scoring scheme of Perlman *et al* (2011). Each feature is based on one drug–drug similarity measure and one disease–disease similarity measure, resulting in 10 features overall. Once the features were constructed, we learned a logistic regression classifier that automatically weighs the different features to yield a classification score.

Performance evaluation and comparison with other methods

To evaluate our classification scheme, we applied it in a 10-fold cross-validation setting. To avoid easy prediction cases, we hid all the associations involved with 10% of the drugs in each

iteration, rather than hiding 10% of the associations. Our method obtained an area under the Receiver-Operating Characteristic Curve (AUC) of 0.9 ± 0.01 (see Materials and methods). In order to validate that the prediction accuracy is not biased by redundant drugs, such as chemically similar drugs or drugs with similar targets, we created three types of non-redundant drug sets by filtering for (i) chemically similar drugs (Tanimoto coefficient >0.7); (ii) drugs with sequence similar drug targets (normalized sequence similarity >0.7); and (iii) drugs sharing at least one target (see Materials and methods). The obtained AUCs remained high: 0.89 ± 0.02 for 427 ± 3 chemically dissimilar drugs, 0.88 ± 0.02 for 290 ± 4 drugs with sequence dissimilar targets and 0.85 ± 0.03 for 99 ± 3 drugs sharing no target. The AUC scores obtained by the algorithm for different Tanimoto coefficient cutoffs are plotted in Supplementary Figure S1, showing comparable performance even at a cutoff of 0.4.

We compared our classification results with two previous methods: (i) the GBA method of Chiang and Butte (2009) and (ii) the CMap approach (Lamb, 2007). Since the GBA method cannot handle drugs for which no associations are known, we compared with it using a modified 10-fold cross-validation setting, in which associations (rather than drugs) are hidden in each iteration. Under this setting, the GBA method obtained false positive rate (FPR) and true positive rate (TPR) scores of 0.13 (± 0.006) and 0.77 (± 0.005), respectively, corresponding to a single point in the Receiver-Operating Characteristic (ROC) space (as GBA does not rank its predictions, a full curve could not be constructed). Supplementary Figure S2 displays the ROC curve of our method in this scenario (AUC=0.913 ± 0.002) and the FPR–TPR point of the GBA method, which falls below our curve.

CMap predicts drug–disease associations by looking for drug response gene expression profiles that anti-correlate with a disease signature. In order to compare with CMap, we downloaded 171 disease genetic signatures from ArrayExpress (Parkinson *et al*, 2009) and mapped them to 239 OMIM diseases by matching a selected set of common UMLS concepts (see Materials and methods). We filtered signature genes with inconclusive regulation direction (i.e., genes that were both upregulated and downregulated across various experiments for the same disease, allowing up to 10% errors; see Materials and methods) and disregarded signatures that did not comply with the restrictions of CMap, including signatures having only upregulated or downregulated genes and signatures with >1000 genes. Overall, 36 signatures survived this filtering, 19 of which mapped to OMIM diseases included in our data set. The intersection of CMap drugs with our set of 593 drugs yielded 69 drugs spanning 149 known associations. The resulting CMap AUC score was small (AUC=0.45; no standard deviation included due to the deterministic nature of the method). Quite strikingly, using our method on this set, we obtain an AUC of 0.92 ± 0.02. In order to compare our performance with CMap over the entire set of 36 signatures, we replaced the phenotypic disease similarity with a signature-based similarity (see Materials and methods). The latter similarity enables comparing CMap predictions for 37 diseases (corresponding to the 36 signatures, with several signatures mapping to more than one OMIM entry) and 71 drugs, spanning 266 associations. The CMap AUC score for this enlarged set, however, remains small (AUC=0.42). In contrast, employing our method replacing the phenotypic disease similarity with the signature-based similarity, we obtained an AUC score of 0.93 ± 0.02. These sets of 19 diseases (using the phenotypic similarity measures) and 37 diseases (using the signature-based similarity measures) were used solely for comparison purpose and are not used in the sequel. We note that Hu and Agarwal (2009) suggested a similar method to CMap, whereby they compared disease and drug expression profiles downloaded from GEO in a similar manner. The authors published a set of predictions, which contains 43 diseases and 45 agents, but the latter could be mapped to merely five diseases and five drugs from our collection (see Materials and methods), preventing a proper comparison.

Analysis of novel predictions

After evaluating our method, we turned to employ it for predicting new drug–disease associations (see Materials and methods). Our first set of predictions consists of new associations for known drugs (i.e., *drug repositioning*) (Figure 2). To prune and validate our predictions, we first used a set of 752 low-confidence associations that were retrieved from only one source of descriptive drug indication and thus were not part of the drug indication gold standard (see Materials and methods). Supplementary Figure S3 displays the hypergeometric enrichment *P*-value of the low-confidence set in our ranked list of predictions, obtained at different score thresholds (the ranked list contains 183 676 possible associations between the 593 drugs and 313 diseases, excluding known associations). The threshold yielding the lowest *P*-value results in a set of 9476 novel drug–disease associations, representing putative drug repositionings for 580 drugs (out of 593 drugs). These associations cover 39% of the low-confidence associations in the data (hypergeometric $P<2\times10^{-177}$).

In order to test whether our predictions are in accordance with current experimental knowledge, we checked the extent to which they appear in current clinical trials. We downloaded drug–disease data from the registry of federally and privately supported clinical trials conducted around the world (http://clinicaltrials.gov/). Overall, we acquired 16 506 unique drug–disease associations that are being investigated in clinical trials (phases I–IV). In all, 2552 of these associations involve drugs and diseases that are present in our data set, spanning 1943 associations that are not part of our gold standard. Using the *P*-value threshold found independently above, we cover 27% of the clinical trial associations (hypergeometric $P<2\times10^{-220}$). The percentage of phase III clinical trial associations predicted by our method is markedly high, covering 38% of the tested associations (hypergeometric $P<6\times10^{-128}$). Table I summarizes the coverage of the predicted associations with respect to the clinical trial associations across the different phases.

To further validate our predictions, we used tissue-specific expression data, motivated by the assumption that if a disease is manifested in a certain tissue then for a drug treating it to have an effect, its targets should be expressed in that tissue. We compared two means of associating a drug with a set of

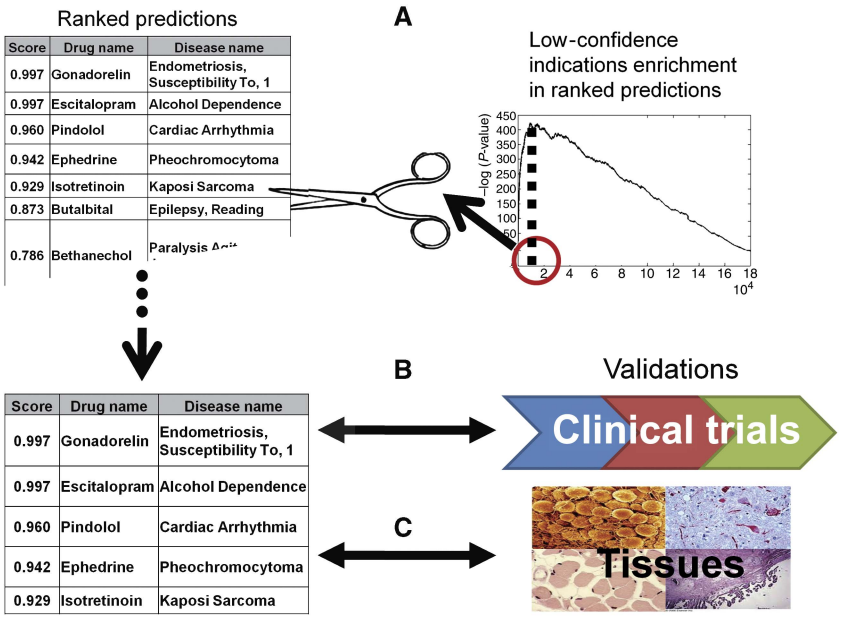

Figure 2 Validation scheme for drug repositioning predictions. We identify a score cutoff that yields the best *P*-value against drug indication originating from single textual indication source (low confidence) (**A**). Applying the cutoff, we validate the selected top ranking predictions against indications under test in clinical trials (**B**) and the co-occurrence of drug targets and indicated diseases in the same tissues (**C**).

Table I Statistics of overlap between the phenotypic-based predictions and drug indications that are under clinical trial

Phases	# of associations in clinical trials		Predicted associations	
	Total	Approved	Coverage (%)	*P*-value
All	2552[a]	609	27	2.0×10^{-220}
I	732	242	32	8.9×10^{-80}
II	1150	324	27	1.1×10^{-94}
III	969	379	38	1.1×10^{-128}
IV	755	311	32	1.9×10^{-69}
Unlisted	719	252	29	1.0×10^{-62}

[a]Unique associations, excluding redundancy between phases.

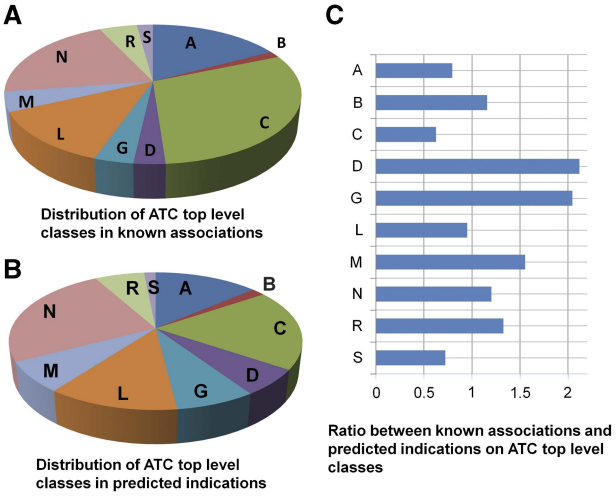

Figure 3 Distributions of Anatomic, Therapeutic and Chemical (ATC) top level classes among drug–disease associations in the gold standard (**A**) and the predicted associations (**B**). The relative ratio between the two distributions for each ATC class is shown in subfigure (**C**). ATC classes include: alimentary tract and metabolism (A), blood and blood forming organs (B), cardiovascular system (C), dermatologicals (D), genito urinary system and sex hormones (G), antineoplastic and immunomodulating agents (L), musculo-skeletal system (M), nervous system (N), respiratory system (R) and sensory organs (S).

tissues: (i) tissues where the drug target is expressed and (ii) tissues affected by diseases that the drug is treating. Using tissue-specific gene expression data by Su *et al* (2004) and literature-based disease–tissue associations of Lage *et al* (2008), we could show that indeed, drug targets are significantly expressed in the same tissues that the indicated diseases affect, supporting our assumption (hypergeometric $P<0.003$; see Materials and methods). Notably, we observed a similar result for our set of predicted associations ($P<0.006$).

The distribution of our predictions across different classes of drugs (according to the Anatomical, Therapeutic and Chemical (ATC) classification system; Skrbo *et al*, 2004) is shown in Figure 3. Notably, we predict significantly more indications for drugs belonging to ATC classes of Dermatologicals and Genito urinary system and sex hormones, which is the result of multiple predictions made for glucocorticoids receptors (e.g., desoximetasone and methylprednisolone) and for sex hormones (testosterone and progesterone), suggesting that these drugs should be further studied for broader indications. The full list of predictions is provided in Supplementary Table S2; and here, we highlight a few interesting examples. For instance, Cabergoline, indicated for Hyperprolactinemia is predicted to treat Migrane. Supporting this prediction, Verhelst *et al* (1999) found in a large study of Cabergoline treatment in Hyperprolactinemia that symptoms of migraine significantly improved in 72% of the patients ; in another smaller scale study, treatment by the drug reduced the number of days of headache per month (Cavestro *et al*, 2006).

As another example, Progesterone is predicted to treat renal cell cancer, non-papillary (npRCC). Indeed, Medroxyprogesterone acetate, a synthetic derivative of progesterone, was suggested as a possible treatment for treating advanced RCC (Izumi *et al*, 2007). Finally, Azathioprine, an immunosuppressive agent, is predicted to treat familial Mediterranean fever, a hereditary inflammatory disorder. A small study has found it successful in completely stopping fever attacks in this disorder (Sayarlioglu *et al*, 2006).

In addition to repositioning current drugs, we examined the utility of our method in predicting indications for 'new' drugs, by inferring disease indications for drugs for which we have no disease indications in our databases. The training was done in the same manner as in the drug repositioning scheme, and the same prediction cutoff value was used. Overall, we attained 3108 predictions which we validated using the same tissue specificity approach described above ($P<0.02$, see Supplementary Figure S4). The full list of predicted indications for new drugs is listed in Supplementary Table S3. In the following, we highlight some of these predictions. The top five predictions for Paget's disease of bone are presented in Supplementary Table S4. We observe that while the two top ranking drugs are chemically similar to drugs indicated for Paget's disease (Risedronate and Alendronate, respectively), the other three bear only moderate chemical similarity to known drugs for Paget's disease (e.g., Salmon Calcitonin) but share a target with another drug for Paget's disease (Tiludronate, targeting tyrosine-protein phosphatase non-receptor type 1). Another interesting finding is the prediction of Cycloleucine for the treatment of Alzheimer's disease (AD). We found that Cycloleucine inhibits methionine adenosyltransferase (MAT) activity, where MAT was abnormally high in brains from patients with AD (Gomes Trolin *et al*, 1998). Furthermore, it was found that Cycloleucine is a potent and selective antagonist of NMDA receptor-mediated responses (Hershkowitz and Rogawski, 1989), a new promising class of chemicals for the treatment of AD (Farlow, 2004). Finally, Hyperforin, St John's wort extract, is predicted to treat hyperthermia. Interestingly, St John's wort extract was found to have anxiolytic effects on stress-induced hyperthermia in mice (Grundmann *et al*, 2006).

Toward personalized medicine—representing diseases by their gene expression profiles

Having shown that our method succeeds in predicting drug–disease associations based on phenotypic descriptions of diseases, we wished to test its application in a personalized medicine scenario, where a patient may be characterized by his/her gene expression signature. To this end, we substituted the disease–disease phenotypic similarity measures with similarity measures based on disease-specific gene signatures. This was done to simulate a scenario where we are given a patient's expression profile and directly query the database with this signature to suggest the best fitting drug for this signature. Using 171 disease signatures from ArrayExpress, we constructed four disease–disease similarity measures (see Materials and methods). Using the same five drug–drug similarities described above and the signature-based disease–disease similarity measures, we were able to predict drug–disease associations for 261 drugs and 114 diseases with an AUC score of 0.92 ± 0.01. The intersection between these 114 diseases and the 313 diseases represented in the phenotypic similarities includes 47 diseases. Comparing the AUC scores obtained on these 47 diseases using both types of similarities, it was encouraging to observe that the signature-based similarities performed identically to the phenotypic ones, while combining the two types of similarities only slightly improved the results (AUC increase of 0.01).

We further used the signatures to provide drug repositioning predictions. Using the same low-confidence associations (i.e., drug–disease associations that were retrieved from only one source of descriptive drug indication) to determine a cutoff, as described for the phenotypic similarity measure, we obtained a list of 2103 novel predictions. These predictions significantly matched low-confidence associations (see Supplementary Figure S5 for the resulting *P*-values, best *P*-value $<1.4\times10^{-101}$) and associations that are currently in clinical trials ($P<1.4\times10^{-30}$, see Supplementary Table S5). They also exhibited significant agreement with the tissue-specific gene expression information ($P<4.1\times10^{-16}$). Importantly, the high predictive power of the gene expression signatures serves as an initial proof-of-concept for the possible future utility of our method also in this personalized medicine setting.

Conclusions

We presented PREDICT—a scheme for similarity-based inference of novel drug indications, capable of handling both approved and novel drugs. Our method attained high rates of specificity and sensitivity in cross-validation (AUC=0.9), surpassing existing methods. Furthermore, our predictions attained significant coverage of drug–disease associations tested in clinical trials and are in good agreement with tissue-specific expression information on the targets of the corresponding drugs, suggesting that they can be regarded as valuable leads for further research.

An important property of our method is that it allows easy integration of additional similarity measures among diseases and drugs. We exploited this property to broaden the applicability of our method by incorporating disease-specific genetic signatures as similarity measures. This will likely to be of significance in the near future, with the accumulation of patient-specific genetic information, such as expression profiles or genotypes, in conjunction with detailed medical records. Our approach may then serve as a basis for patient tailored predictions.

Materials and methods

Data sets

Drug targets, drug indications and canonical simplified molecular input line entry specification (SMILES) (Weininger, 1988) of the drugs were extracted from DrugBank (Wishart *et al*, 2008). Additional drug targets were obtained from the DCDB (Liu *et al*, 2010), Matador (Gunther *et al*, 2008) and KEGG DRUG (Kanehisa and Goto, 2000) databases. FDA drug labels and additional drug indications were downloaded from the DailyMed site (http://dailymed.nlm.nih.gov)

and from http://drugs.com. Drug side effects were obtained from SIDER (Kuhn *et al*, 2010). Human PPIs were compiled from experimental and literature curated data (Xenarios *et al*, 2002; Rual *et al*, 2005; Stelzl *et al*, 2005; Ewing *et al*, 2007; Breitkreutz *et al*, 2008). Protein sequences and GO annotations (Ashburner *et al*, 2000) were downloaded from UniProt (Jain *et al*, 2009). Clinical trials data were downloaded from a registry of federally and privately supported clinical trials (http://clinicaltrials.gov/). Disease signatures were downloaded from the Gene Expression Atlas of ArrayExpress (Parkinson *et al*, 2009).

Assembling drug–disease associations

We assembled associations between diseases listed in the OMIM (Hamosh *et al*, 2002) and their indicated drugs, registered in DrugBank (Wishart *et al*, 2008). OMIM is a comprehensive disease phenotype database, encompassing thousands of phenotypic descriptions of diseases and disorders, including single-gene as well as complex multiple-gene disorders, allowing for the construction of phenotypic similarity measures.

In order to handle variations in disease names between OMIM and drug indication sources, we mapped disease names to UMLS concepts (Bodenreider, 2004), serving as a standardized terminology. We began by mapping OMIM disease names to UMLS concepts, exploiting the integrated curated mapping between OMIM and UMLS concepts and augmented the mapping by using the MetaMap tool (Aronson, 2001). The MetaMap tool uses symbolic, natural language processing and computational linguistic techniques to map biomedical text to the UMLS metathesaurus, and was previously reported to exceed human mapping (to UMLS concepts) capabilities (Pratt and Yetisgen-Yildiz, 2003). We permitted only disease-related UMLS concept types (e.g., 'Disease or Syndrome', 'Anatomical Abnormality' or 'Neoplastic Process') and filtered ambiguous or generic UMLS concepts (e.g., 'Infections' or 'Syndrome'). We further augmented the MetaMap mapped UMLS concepts with the rich synonymy relationships embedded in the UMLS. Supplementary Table S6 lists the mapping between OMIM disease names and UMLS concepts.

The associations between drugs and UMLS disease concepts were integrated from four different sources using three different methods: (i) direct mapping to drugs, exploiting embedded UMLS links between concepts and drugs; (ii) drug–condition associations downloaded from http://drugs.com, where conditions were mapped to UMLS concepts using MetaMap; and (iii) indication-based mapping. For the latter, we extracted UMLS concepts using the MetaMap tool from textual drug indications downloaded from FDA package inserts (available in the DailyMed website, http://dailymed.nlm.nih.gov) and DrugBank. In addition, we manually added 44 associations occurring in phase IV (post-marketing) clinical trials.

In order to deal with variations in drug names, we used generic and synonymous drug names; if no match was found, we used also brand names retrieved from DrugBank. We removed UMLS concepts matching hard to treat congenital anomalies such as congenital malformations, chromosomal abnormalities and inborn errors of metabolism using the ICD-10. We filtered disease names classified under the ICD-10 chapter XVII (Congenital malformations, deformations and chromosomal abnormalities (Q00–Q99)) and metabolic disorders (E70–E90) by mapping them to UMLS concepts using MetaMap followed by manual curation of the list, resulting in the removal of 263 OMIM diseases. Finally, performing a manual curation of the extracted UMLS concepts from textual description of drug indications, we observed that they are more prone to false positives. We thus required that associations extracted from drug indications appear also in at least one more source.

Similarity measures

We defined and computed five drug–drug similarity measures and two disease–disease similarity measures. Three of the drug–drug similarities (3–5) are gene related, based on the drug targets downloaded from DrugBank. For drugs associated with more than one gene, maximal similarities between the associated genes were averaged (averaging over the contribution of each member in a drug pair for symmetry). All similarity measures were normalized to be in the range (0, 1).

We used the following drug–drug similarity measures:

(1) *Chemical based:* Canonical SMILES (Weininger, 1988) of the drug molecules were downloaded from DrugBank (Wishart *et al*, 2008). Hashed fingerprints were computed using the Chemical Development Kit with default parameters (Steinbeck *et al*, 2006). The similarity score between two drugs is computed based on their fingerprints according to the two-dimensional Tanimoto score (Tanimoto, 1957), which is equivalent to the Jaccard score (Jaccard, 1908) of their fingerprints, that is, the size of the intersection over the union when viewing each fingerprint as specifying a set of elements.

(2) *Side effect based:* Drug side effects were obtained from SIDER (Kuhn *et al*, 2010), an online database containing drug side effects associations extracted from package inserts using text mining methods. We augmented this list by side effect predictions for drugs that are not included in SIDER based on their chemical properties (Atias and Sharan, 2011). Following this latter work, we defined the similarity between drugs according to the Jaccard score between either their known side effects or top 10 predicted side effects in case they are unknown.

(3) *Sequence based:* Based on a Smith–Waterman sequence alignment score (Smith *et al*, 1985) between the corresponding drug-related genes. Following the normalization suggested in Bleakley and Yamanishi (2009), we divide the Smith–Waterman score by the geometric mean of the scores obtained from aligning each sequence against itself.

(4) *Closeness in a PPI network:* The distances between each pair of drug-related genes were calculated on their corresponding proteins using an all-pairs shortest paths algorithm on the Human PPI network. Distances were transformed to similarity values using the formula described in Perlman *et al* (2011):

$$S(p, p') = Ae^{-bD(p, p')} \quad (1)$$

where $S(p,p')$ is the computed similarity value between two proteins, $D(p,p')$ is the shortest path between these proteins in the PPI network and A, b were chosen according to Perlman *et al* (2011) to be $0.9 \times e$ and 1, respectively. Self similarity was assigned a value of 1.

(5) *GO based:* Semantic similarity scores between drug-related genes were calculated according to Resnik (1999), using the csbl.go R package (Ovaska *et al*, 2008) selecting the option to use all three ontologies.

For diseases, we employed two sets of measures, depending on whether we wished to exploit phenotypic (1–2) or gene signature information (3–6). As in drugs similarities, maximal values between the two lists of associated genes were averaged (taking into account both sides for symmetry). The disease–disease similarity measures we used include:

(1) *Phenotype based:* We used the phenotypic similarity constructed by van Driel *et al* (2006). The phenotypic similarity was constructed by identifying similarity between MeSH terms (Lipscomb, 2000) appearing in the medical description of diseases from the OMIM database (Hamosh *et al*, 2002).

(2) *Semantic phenotypic similarity:* We used the hierarchical structure of the HPO (Robinson and Mundlos, 2010) together with the mapping provided by HPO between ontology nodes and OMIM diseases to construct a semantic similarity score based on Resnik (1999). Applying this semantic similarity on the HPO data was previously shown to provide consistent clustering of diseases (Robinson *et al*, 2008).

(3) *Genetic based:* Given genetic signatures of diseases obtained from gene expression experiments, we used a Jaccard score between every pair of signatures, taking into account the direction of the response of each gene. That is, the total number of mutual upregulated genes and mutual downregulated genes over the unified list of all genes. Signature genes with inconsistent regulation directionality for the same disease across various experiments (i.e., registered as both upregulated and downregulated across various experiments for the same disease) were filtered, allowing for up to 10% expression measurement errors.

(4–6) *Signature sequence based, signature PPI based and signature GO based:* Similar to the drug–drug sequence-, PPI- and GO-based

similarity, respectively, we used these similarities between genes participating in the two signatures.

Combining measures to classification features

The classification features that we used were constructed from association scores calculated on pairs of drug–drug and disease–disease similarity measures, resulting in 10 features when using the phenotypic disease similarities and 20 features when using the signature-based similarities. For a given similarity measure pair (i.e., feature), the score of a given drug–disease association (d_r, d_i) is calculated by considering the similarity, according to the given pair, of all known drug–disease associations to this association. The computation is done as follows: First, for each known associations (d_r', d_i') we compute the drug–drug similarity $S(d_r, d_r')$ and the disease–disease similarity $S(d_i, d_i')$. Next, we follow the method of Perlman *et al* (2011) to combine the two similarities to a single score by computing their weighted geometric mean. Here, we chose to use a simple (unweighted) geometric mean as the resulting AUC score was robust to the weighting parameter under a wide range (AUC difference <0.01, data not shown). Thus:

$$\text{Score}(d_r, d_i) = \max_{d_r', d_i' \neq d_r, d_i} \sqrt{S(d_r, d_r') \times S(d_i, d_i')} \qquad (2)$$

Prediction assessment

We used a 10-fold cross-validation scheme to evaluate the accuracy of our prediction algorithm. To concentrate on 'hard' cases, we hid 10% of the drugs in each iteration rather than 10% of the associations. The training set used for the cross-validation included 1933 true drug–disease associations and a randomly generated set of drug–disease pairs (not part of the positive set), twice as large as the positive set. To obtain robust AUC score estimates, we performed 100 independent cross-validation runs, in each of which a different random partition of the training set to 10 parts was used; we then averaged the resulting AUC scores. We note that taking a negative set of the same size as the positive set, three-fold larger or even 50 times as large had a negilible effect on the resulting AUC score (<0.01). In order to test the effect of redundant drugs on prediction accuracy, we created non-redundant sets of drugs filtered by three different methods: (i) chemical similarity above a Tanimoto coefficient ranging from 0.85 to 0.4; (ii) normalized sequence similarity between their targets ranging from 1 to 0.7; and (iii) target sharing. To this end, we iteratively selected the most similar pair and randomly removed one of the pair's drugs. We repeated this procedure 50 times for each similarity threshold to construct different non-redundant sets and averaged over the resulting AUC score (reporting also the AUC standard deviation).

To evaluate the benefit of using a feature selection scheme, we employed both forward feature selection and backward feature elimination in a cross-validation setting to select the best-performing features. We found that the difference between the best feature set according to each feature selection method and using all the features was negligible (AUC difference <0.005). We thus retained the entire set of features.

We used the MATLAB implementation of the logistic regression classifier (glmfit function with binomial distribution and logit linkage) for the prediction task.

Comparison with other methods

We implemented the GBA method of Chiang and Butte (2009) and applied it to the set of drugs and diseases appearing in our data. Since it cannot be tested using a cross-validation scheme that involves the removal of entire drugs, we tested it by removing associations instead.

In order to compare our results with the CMap, we obtained disease gene signatures from the Gene Expression Atlas of ArrayExpress (Parkinson *et al*, 2009). We mapped the ArrayExpress disease titles to UMLS concepts using the same filtering procedures described above. These concepts were then matched to OMIM disease mapped UMLS concepts. We queried the online web tool of CMap with the signatures and followed CMap suggestion to select drugs obtaining negative enrichment (the signature has opposite effect to the drug expression profile) and non-null *P*-value. An ROC curve was obtained by choosing different *P*-values as cutoffs.

Hu and Agarwal (2009) suggested a similar method to CMap, whereby they predicted drug–disease associations based on disease and drug profiles downloaded from GEO. The authors' published set of predictions contain 43 profiles tagged as diseases or infections and 45 profiles tagged as agents. We mapped disease names to OMIM identifiers using the same procedure used for constructing the gold standard, resulting in 17 mapped OMIM diseases. Mapping drug names, we matched generic, synonymous and brand names from DrugBank followed by additional manual mapping of unmapped names, obtaining 13 drugs successfully mapped to DrugBank entries. However, only five diseases and five drugs, connected by 10 predicted associations (out of the 17 diseases and 13 drugs) intersected our collection, preventing a proper comparison with this method.

Novel predictions

To predict novel indications for drugs, we used a training set that included all the known associations and a two-fold larger randomly generated set of drug–disease pairs that are not known to be associated (i.e., associations that do not appear in our drug indication gold standard). We applied the trained classifier to all remaining drug–disease pairs to form our prediction set. We repeated the analysis with another randomly picked negative set, distinct from the first one, to assign prediction scores also to the random negative set that we initially used for training. Overall, we obtained classification scores for all 183 676 unknown drug–disease pairs.

We validated the predictions in two ways: (i) by comparing predicted drug–disease associations with drug–disease associations being investigated in clinical trials (phases I–IV) and (ii) by testing their agreement with tissue-specific gene expression information—that is, by comparing the tissues where the drug targets are expressed with tissues associated with the diseases predicted to be treated by the drug. All *P*-values were computed using a hypergeometric test.

For the clinical trial validation, we downloaded phases I–IV clinical trials from http://clinicaltrials.gov/ and mapped the disease names to UMLS concepts using the MetaMap tool and drug names to DrugBank using the same procedures described above. For the tissue-related validation, we assembled two drug–tissue association matrices. The first matrix was constructed based on the tissue-specific gene expression of Su *et al* (2004). A drug was declared to affect a tissue if at least one of its targets is expressed in that tissue over a certain tunable threshold (see below). The second drug–tissue matrix was assembled by exploiting the disease–tissue associations computed by Lage *et al* (2008). Lage *et al* scored disease–tissue associations according to co-mentioning of diseases and tissues across PubMed (Korbel *et al*, 2005).Thus, a drug was associated with a tissue if at least one of its indicated diseases was associated with that tissue. We manually matched the tissue names specified by Su *et al*, to tissue names specified by Lage *et al*, resulting in 67 common tissues.

Both the gene expression–tissue associations and the disease–tissue associations require thresholds to define true associations. In order to decouple the prediction task from tuning these thresholds, we randomly split the gold standard set of drugs into two disjoint groups (two thirds and a third), where we provide predictions only based on the larger group and use the smaller group for tuning the thresholds. We used the Wilcoxon rank sum test to test whether the drug targets are significantly expressed in tissues related to the diseases these drugs are predicted to treat. By the learned thresholds, genes with an expression value <380 were declared unassociated with a certain tissue, interestingly corresponding to the average expression in the entire data, and diseases with co-occurrence score <18 in a certain tissue were declared unrelated to that tissue. During learning, the Wilcoxon rank sum test yielded a *P*-value of 0.003 on the smaller group of known drug indications. Using the same tuned thresholds on the predictions made using the second group, we obtained a significant *P*-value of 0.006.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Nir Atias for introducing us to MetaMap and UMLS. AG was partially funded by the Edmond J Safra bioinformatics program. RS and ER were partially supported by a Bikura grant from the Israel Science Foundation.

Author contributions: AG and RS conceived the paper; AG perfomed the experiments and analyzed the data; AG and GYS performed manual verification of novel predictions; AG, ER and RS wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 17–21

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

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

Bleakley K, Yamanishi Y (2009) Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* **25:** 2397–2403

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

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, Dolinski K, Tyers M (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36:** D637–D640

Cavestro C, Rosatello A, Marino MP, Micca G, Asteggiano G (2006) High prolactin levels as a worsening factor for migraine. *J Headache Pain* **7:** 83–89

Chiang AP, Butte AJ (2009) Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther* **86:** 507–510

DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. *J Health Econ* **22:** 151–185

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y *et al* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3:** 89

Farlow MR (2004) NMDA receptor antagonists. A new therapeutic approach for Alzheimer's disease. *Geriatrics* **59:** 22–27

Gomes Trolin C, Ekblom J, Oreland L (1998) Regulation of methionine adenosyltransferase catalytic activity and messenger RNA in SH-SY5Y human neuroblastoma cells. *Biochem Pharmacol* **55:** 567–571

Grundmann O, Kelber O, Butterweck V (2006) Effects of St. John's wort extract and single constituents on stress-induced hyperthermia in mice. *Planta Med* **72:** 1366–1371

Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* **36:** D919–D922

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **30:** 52–55

Hansen NT, Brunak S, Altman RB (2009) Generating genome-scale candidate gene lists for pharmacogenomics. *Clin Pharmacol Ther* **86:** 183–189

Hershkowitz N, Rogawski MA (1989) Cycloleucine blocks NMDA responses in cultured hippocampal neurones under voltage clamp: antagonism at the strychnine-insensitive glycine receptor. *Br J Pharmacol* **98:** 1005–1013

Hu G, Agarwal P (2009) Human disease-drug network based on genomic expression profiles. *PLoS One* **4:** e6536

Izumi K, Kanno H, Umemoto S, Hasumi H, Osada Y, Otai J, Mikata K, Tsuchiya F, Nagashima Y (2007) [A case of recurrent renal cell carcinoma which recurred after fourth surgical resection and survived for about 2 years by medroxyporgesterone acetate administration]. *Hinyokika Kiyo* **53:** 623–626

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

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10:** 136

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30

Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE (2009) Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput Biol* **5:** e1000423

Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3:** e134

Kotelnikova E, Yuryev A, Mazo I, Daraselia N (2010) Computational approaches for drug repositioning and combination therapy design. *J Bioinform Comput Biol* **8:** 593–606

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* **6:** 343

Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci USA* **105:** 20870–20875

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

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313:** 1929–1935

Li J, Zhu X, Chen JY (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput Biol* **5:** e1000450

Lipscomb CE (2000) Medical subject headings (MeSH). *Bull Med Libr Assoc* **88:** 265–266

Liu Y, Hu B, Fu C, Chen X (2010) DCDB: drug combination database. *Bioinformatics* **26:** 587–588

Ovaska K, Laakso M, Hautaniemi S (2008) Fast gene ontology based clustering for microarray experiments. *BioData Min* **1:** 11

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ *et al* (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* **37:** D868–D872

Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R (2011) Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol* **18:** 133–145

Pratt W, Yetisgen-Yildiz M (2003) A study of biomedical concept identification: MetaMap vs. people. *AMIA Annu Symp Proc* 529–533

Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res* **11:** 95–130

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83:** 610–615

Robinson PN, Mundlos S (2010) The human phenotype ontology. *Clin Genet* **77:** 525–534

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S *et al* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437:** 1173–1178

Sayarlioglu H, Erkoc R, Sayarlioglu M, Dogan E, Soyoral Y (2006) Successful treatment of nephrotic syndrome due to FMF amyloidosis with azathioprine: report of three Turkish cases. *Rheumatol Int* **27:** 197–199

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

Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res* **13:** 645–656

Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (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, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B *et al* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957–968

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101:** 6062–6067

Tanimoto TT (1957) IBM Internal Report 17th Nov

Terstappen GC, Reggiani A (2001) In silico research in drug discovery. *Trends Pharmacol Sci* **22:** 23–26

van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. *Eur J Hum Genet* **14:** 535–542

Verhelst J, Abs R, Maiter D, van den Bruel A, Vandeweghe M, Velkeniers B, Mockel J, Lamberigts G, Petrossians P, Coremans P, Mahler C, Stevenaert A, Verlooy J, Raftopoulos C, Beckers A (1999) Cabergoline in the treatment of hyperprolactinemia: a study in 455 patients. *J Clin Endocrinol Metab* **84:** 2518–2522

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

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36:** D901–D906

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30:** 303–305