Bioinformatics, 36, 2020, i516–i524
doi: 10.1093/bioinformatics/btaa439
ISMB 2020

Network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets

Midori Iida, Michio Iwata and Yoshihiro Yamanishi*

Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

*To whom correspondence should be addressed.

Abstract

Motivation: Disease states are distinguished from each other in terms of differing clinical phenotypes, but characteristic molecular features are often common to various diseases. Similarities between diseases can be explained by characteristic gene expression patterns. However, most disease–disease relationships remain uncharacterized.
Results: In this study, we proposed a novel approach for network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets. We performed large-scale analyses of omics data and molecular interaction networks for 79 diseases, including adrenoleukodystrophy, leukaemia, Alzheimer's disease, asthma, atopic dermatitis, breast cancer, cystic fibrosis and inflammatory bowel disease. We quantified disease–disease similarities based on proximities of abnormally expressed genes in various molecular networks, and showed that similarities between diseases could be explained by characteristic molecular network topologies. Furthermore, we developed a kernel matrix regression algorithm to predict the commonalities of drugs and therapeutic targets among diseases. Our comprehensive prediction strategy indicated many new associations among phenotypically diverse diseases.
Contact: yamani@bio.kyutech.ac.jp
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Disease states are distinguished from each other in terms of differing clinical phenotypes, but recent medical studies have uncovered commonalities between different diseases. For example, whereas erectile dysfunction and pulmonary hypertension are phenotypically different diseases, they share the therapeutic target protein phosphodiesterase 5 (PDE5; Boolell *et al.*, 1996; Weimann *et al.*, 2000). The approved drug sildenafil (Viagra) is hence prescribed for both of these diseases, suggesting that disease pairs with shared therapeutic targets can inform drug discovery. However, disease phenotypes follow highly complex interactions between numerous biomolecules. Defects in PDE5 are known consequences of post-transcriptional regulation (Boolell *et al.*, 1996; Weimann *et al.*, 2000). Therefore, it is important to investigate disease–disease relationships in terms of molecular interaction networks.

A network-based approach is useful for analyzing drug–drug, disease–disease and drug–disease associations. As examples of drug–disease associations, target proteins of approved drugs reportedly locate neighbourhoods of disease-related proteins in such networks (Yıldırım *et al.*, 2007). Subsequently, a drug–disease proximity measure was proposed to quantify the therapeutic effects of drugs in a network of disease-related genes. This drug–disease proximity measure identified novel uses for existing drugs (Guney *et al.*, 2016). Moreover, to identify disease–disease associations, subnetworks (modules) were identified in the network using disease-related genes. These modules enable determinations of pathological relationships, such as co-expression patterns, symptom similarities and shared comorbidities among diseases. These efforts also confirmed that phenotypically related diseases have similar disease-causing genes, whereas primary disease-related genes are not always shared. Although types of molecular interactions were not distinguished in the previous works, the importance of network-based approaches to identifying relationships between diseases was suggested.

Despite the promise of network-based approaches, disease-causing genes did not always correspond with therapeutic targets of diseases. Specifically, disease–gene interactions were elucidated using the information on genetic disorders (Guney *et al.*, 2016; Menche *et al.*, 2015; Yıldırım *et al.*, 2007), in which Online Mendelian Inheritance in Man (OMIM; Amberger and Hamosh, 2017), Genome Wide Association Study catalog, the UniProt Knowledgebase (UniProtKB; Mottaz *et al.*, 2008) and the Phenotype–Genotype Integrator (PheGenI; Ramos *et al.*, 2014) were used. These databases contain huge numbers of gene defects and related diseases. However, most disease phenotypes cannot be explained by gene defects alone, and disease pathogenesis is better defined by disruptions of coordinated gene expression systems in cells.

In this study, we proposed a novel approach for network-based characterization of disease–disease relationships in terms of drugs and therapeutic targets. We performed large-scale diseasome analyses of omics data and molecular interaction networks for 79 diseases, including adrenoleukodystrophy, leukaemia, Alzheimer's disease, asthma, atopic dermatitis, breast cancer, cystic fibrosis and inflammatory bowel disease. We quantified disease–disease similarities based on the proximities of abnormally expressed genes in various molecular networks, and suggested that similarities between diseases could be explained by characteristic molecular network topologies. Furthermore, we developed a kernel matrix regression

algorithm to predict the commonalities of drugs and therapeutic targets among diseases. Our comprehensive prediction suggests many new associations among some phenotypically different diseases.

2 Materials and methods

2.1 Disease-specific gene expression profiles

Disease-specific gene expression profiles were constructed based on gene expression profiles in CRowd Extracted Expression of Differential Signatures (CREEDS; Wang *et al.*, 2016). Initially, we retrieved gene expression profiles for patients with 695 diseases with assigned disease ontology IDs (DOIDs; Kibbe *et al.*, 2015) annotated as manual disease signatures v1.0. The gene expression profiles comprised scores that were calculated using the characteristic direction method (Clark *et al.*, 2014), which compares gene expression levels in diseased tissues with those in control tissues. According to a previous study (Iwata *et al.*, 2019), we converted these DOIDs into the Kyoto Encyclopaedia of Genes and Genomes (KEGG) DISEASE (Kanehisa *et al.*, 2010) IDs using medical subject headings terms or the OMIM database (Hamosh, 2002). Genes with non-zero expression scores were considered disease-associated genes. In total, we constructed 14 804-dimensional gene expression profiles for 79 diseases.

2.2 Molecular interaction networks

To characterize disease–disease relationships in molecular interaction networks, we constructed several types of networks of protein–protein interactions (human interactome). We curated protein–protein interactions from the same databases and denoted these as 'Y2H interactions', 'protein complexes', 'kinase–substrate pairs', 'metabolic enzyme-coupled interactions' and 'signalling interactions' according to the types of interactions. These interaction types are described in detail below:

- Y2H interactions

We combined several yeast-two-hybrid high-throughput datasets from the literature (Rolland *et al.*, 2014; Rual *et al.*, 2005; Stelzl *et al.*, 2005; Venkatesan *et al.*, 2009; Yang *et al.*, 2016; Yu *et al.*, 2011) with a yeast two-hybrid database (HuRI). These data sources together yielded 57 942 interactions between 9441 proteins.

- Protein complexes

Protein complexes are single molecular units that integrate multiple gene products. The comprehensive resource of mammalian protein complexes (CORUM) database (Giurgiu *et al.*, 2019) is a collection of mammalian complexes that were characterized in co-immunoprecipitation, co-sedimentation and ion-exchange chromatography analyses. In total, CORUM yields 2837 protein complexes with 3067 proteins connected by 38 876 interactions.

- Kinase–substrate pairs

Protein kinases are crucial regulators of many biological processes, and principally act through specific signal transduction pathways. PhosphositePlus (Hornbeck *et al.*, 2012) provided a network of peptides that can be bound by kinases, yielding a total of 5424 interactions between 2424 proteins.

- Metabolic enzyme-coupled interactions

When two enzymes share adjacent reactions, they are assumed to be coupled. We obtained enzyme-coupled interactions from the KEGG RPAIR database (Shimizu *et al.*, 2008). The database contains 70 033 metabolic interactions between 1765 enzymes.

- Signalling interactions

Signalink 2.0 (Fazekas *et al.*, 2013) is a multi-layered database of signalling pathways and their regulators, such as scaffold and endocytotic proteins, and modifier enzymes, such as phosphatases, ubiquitin ligases, and transcriptional and post-transcriptional regulators of all components in *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. We downloaded all layers of human associations, including manually curated pathways (pathway members), endocytotic and scaffold proteins (pathway regulators), pathway protein modifying enzymes (post-translational modifications), possible first neighbours of pathway proteins with direction (directed protein–protein interactions), transcription factor (TF)–TF binding site (promoter) interactions (transcriptional regulators), miRNA–mRNA interactions and transcriptional regulation of miRNAs (post-transcriptional regulators), and undirected protein–protein interactions from small-scale and high-throughput databases (further interactions). Because the genes were listed with gene symbols in the database, these were converted into ensemble IDs. After conversion, we identified 49 590 interactions between 2155 proteins.

- Mixture

In a previous study, molecular interaction networks were constructed by combining different types of protein–protein interactions into a single one (Menche *et al.*, 2015). In further experiments, we integrated all interactions from these databases and found 219 192 interactions between 12 648 proteins. This interactome data is denoted 'mixture' in this paper.

Table 1 summarizes the number of proteins, the number of interactions, the network density and the average number of edges for each node (average degree). Maximum interactions are available from 'metabolic enzyme-coupled interactions', followed by 'Y2H interactions' and 'signalling interactions'. The maximum average degrees are available from 'metabolic enzyme-coupled interactions', suggesting that each protein in the network has a relatively large number of interactions. In contrast, the network based on 'kinase–substrate pairs' was relatively small. The networks 'Y2H interactions', 'protein complexes' and 'kinase–substrate pairs', have sparse network densities.

2.3 Therapeutic targets and drugs of diseases

We identified 1523 disease–target associations involving 100 diseases and 579 therapeutic target proteins from medical books (Papadakis *et al.*, 2013) and the KEGG DISEASE (Kanehisa *et al.*, 2010) database. We also obtained 5606 drug–disease associations involving 2266 drugs and 461 diseases from medical books (Papadakis *et al.*, 2013) and the KEGG DISEASE (Kanehisa *et al.*, 2010) database.

3 Methods

3.1 Disease similarity based on proximity in networks

To consider disease similarities based on their molecular networks, proximities between diseases must be evaluated in the network. In a previous study (Guney *et al.*, 2016), proximity was evaluated using a separation distance measure that was defined by path lengths between disease-related proteins.

Herein, we denoted the set of related proteins for diseases A and B as S_A and S_B, respectively. To define the shortest path lengths between these sets of disease-related proteins (S_A and S_B), we calculated degrees of dispersion between disease-related protein sets as follows:

$$\text{dispersion}(S_A, S_B) = \frac{||S_A||d_c(S_B, S_A) + ||S_B||d_c(S_A, S_B)}{||S_A|| + ||S_B||},$$

where $||S_A||$ is the number of related proteins for disease A, $||S_B||$ is the number of related proteins for disease B, and $d_c(S_A, S_B)$ is the closest measure between diseases A and B. The closest measure is defined as

$$d_c(S_A, S_B) = \frac{1}{||A|| + ||B||}\left(\sum_{a\in A}\min_{b\in B}d(a,b) + \sum_{b\in B}\min_{a\in A}d(a,b)\right).$$

We then defined the shortest path length d_s between the set of disease-related proteins S_A and S_B as follows:

Table 1. Summary of molecular interaction networks

Types of interactions	Database	Number of proteins	Number of interactions	Network density	Average degree
Y2H interactions	HuRI	9441	57 942	0.001	12
Protein complexes	CORUM	3067	38 876	0.006	21
Kinase–substrate pairs	PhosphsitePlus	2424	5424	0.002	4
Metabolic enzyme-coupled interactions	KEGG repair	1765	70 033	0.045	79
Signalling interactions	Signalink	2155	49 590	0.021	46
Mixture	–	12 648	219 192	0.003	34

The best score for each method is highlighted in bold.

$$d_s(S_A, S_B) = \text{dispersion}(S_A, S_B) - \frac{d'_c(S_A, S_A) + d'_c(S_B, S_B)}{2},$$

where d'_c is the modified closest measure in which the shortest path length from the node to itself is set as infinite.

To evaluate the distances between diseases, the shortest path length was normalized using the distribution of randomized shortest path lengths. It is referred to as proximity measure (Guney *et al.*, 2016). In this study, we obtained the network-based disease similarity from the proximity measure by the sign inversion and scaling in the range from 0 to 1. For example, the reference length for the pair of diseases A and B was repeatedly calculated using two randomly selected sets of proteins, with certain criteria for random selection. The number of proteins in randomly selected sets corresponded with those in diseases A or B. Moreover, degrees of each protein basically corresponded with those of each protein in diseases A or B. After 100 repetitions, we calculated the mean $\mu(S_A, S_B)$ and the standard deviation $\sigma(S_A, S_B)$ and defined the normalized proximity $z(S_A, S_B)$ as follows:

$$z(S_A, S_B) = \frac{d(S_A, S_B) - \mu(S_A, S_B)}{\sigma(S_A, S_B)}.$$

Finally, we obtained network-based disease similarity scores by scaling the normalized proximity measure between 0 and 1.

Note that, due to the scale-free nature of the human interactome, few nodes have high degrees. Thus, to avoid repeatedly choosing the same nodes during random selection, we applied a binning approach in which nodes within certain degree intervals were grouped so that at least 100 nodes were included in the bin.

3.2 Algorithm for predicting commonalities among diseases

Given two diseases, we consider predicting the commonality between the diseases. In this study, we evaluated the disease–disease commonality by the commonality of drugs or therapeutic targets between different diseases. For example, if two different diseases share at least one drug, the two diseases are considered to have commonality in terms of drugs. Likewise, if two different diseases share at least one therapeutic target, the two diseases are considered to have commonality in terms of therapeutic targets. We attempt to predict if different diseases would share the same therapeutic targets, and we also attempt to predict if different diseases would share the same drugs.

Suppose that we have an explanatory random variable $x \in R^d$ and a response random variable $y \in \{0, 1\}^l$ for a set of diseases. We considered that the information on the omics data is available for all N diseases, but the information on therapeutic targets/drugs is available for the first n diseases and not for the remaining $(N-n)$ diseases. Accordingly, we refer to the first n diseases as the *training set*, and the remaining $(N-n)$ diseases as the *prediction set*.

Let k and g be symmetric positive definite kernels for x and y, respectively. When we compute the kernel matrix for the explanatory variable x, we obtain an $N \times N$ kernel matrix K, where $(K)_{ij} = k(x_i, x_j)\,(1 \le i, j \le N)$, x_i indicates the i-th disease with omics data, and N is the number of all diseases. In contrast, when we compute the kernel matrix for the response variable y, we obtain the $N \times N$ kernel matrix G, where $(G)_{ij} = g(y_i, y_j)\,(1 \le i, j \le n)$, y_i indicates the i-th disease with the information on therapeutic targets/drugs, and n is the number of available diseases $(n < N)$. Note that G contains missing values for all entries $(G)_{ij}$ with $\max(i, j) > n$. In this study, K corresponds to a similarity matrix of diseases with omics data, and G corresponds to an adjacency matrix in which each element indicate if two diseases share the same therapeutic targets/drugs or not.

To estimate the missing part of G using full Gram matrix K, we considered a form of correlation between the two kernels. We express each kernel matrix by splitting the matrix into four parts. $K_{tt}(\text{resp. } G_{tt})$ denotes the $n \times n$ kernel matrix for the *training set* versus itself, $K_{pt}(\text{resp. } G_{pt})$ denotes the $(N-n) \times n$ kernel matrix for the *prediction set* versus the *training set*, and $K_{pp}(resp.\ G_{pp})$ denotes the $(N-n) \times (N-n)$ kernel matrix for the *prediction set* versus itself as follows:

$$K = \begin{pmatrix} K_{tt} & K_{pt}^\top \\ K_{pt} & K_{pp} \end{pmatrix}, \quad G = \begin{pmatrix} G_{tt} & G_{pt}^\top \\ G_{pt} & G_{pp} \end{pmatrix}.$$

Note that K_{pt} and K_{pp} are known and G_{pt} and G_{pp} are unknown. The objective is to predict G_{pt} and G_{pp} from K and G_{tt}. Here, we describe two approaches to solve the problem of kernel matrix completion.

(i) Unsupervised approach with similarity values

The most straightforward approach is to directly use K_{pt} and K_{pp} for G_{pt} and G_{pp}, respectively. The approach is referred to as unsupervised approach. We directly used disease similarity scores as prediction scores; therefore, similar diseases are predicted to share drugs and therapeutic targets. This framework is similar to a previous work on disease–disease relationship analysis (Guney *et al.*, 2016).

(ii) Supervised approach with kernel matrix regression (KMR)

In reality, the commonality information on therapeutic targets and drugs are partially known for a limited number of diseases. To incorporate the pre-knowledge on disease commonality in the supervised learning framework, we propose a variant of the regression model based on the underlying features in the reproducing kernel Hilbert space.

The ordinary regression model between the explanatory variable $x \in R^d$ and the response variable $y \in R$ can be formulated as follows:

$$y = f(x) + \epsilon, \tag{1}$$

where $f : R^d \to R$ and ϵ is a noise term. By analogy, we regarded $(x, x') \in R^d \times R^d$ as an explanatory variable and $g(y, y') \in R$ as a response variable in the present context. Assuming the underlying feature $u(x) \in R^m$ in the reproducing kernel Hilbert space, we formulated a variant of the regression model as follows:

$$g(y, y') = f(x, x') + \epsilon = u(x)^\top u(x') + \epsilon, \tag{2}$$

where $f : R^d \times R^d \to R$. We refer to this model as a KMR model. We note that imposing f to be of the form $f(x, x') = u(x)^\top u(x')$ for some feature $u : R^d \to R^m$ ensures that the regression function is

positive and definite and the number of dimensions m of the feature u is allowed to be infinite.

Following kernel methods, we consider features in the reproducing kernel Hilbert space of kernel K that possess an expansion of the form:

$$u(x) = \sum_{j=1}^{n} k(x, x_j) w_j, \tag{3}$$

where $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)^\top$ is a weight vector and n is the number of diseases in the *training set*. When m different features are considered, we express them using the feature vector u as $u(x) = (u^{(1)}(x), u^{(2)}(x), \cdots, u^{(m)}(x))^\top$.

To represent the set of features for all diseases, we defined the feature score matrices $U_t(x) = [u(x_1), u(x_2), \cdots, u(x_n)]^\top$ for the *training set* and $U_p(x) = [u(x_{n+1}), u(x_{n+2}), \cdots, u(x_N)]^\top$ for the *prediction set*.

In the matrix form, we computed feature score matrices as $U_t = K_{tt}W$ for the *training set* and $U_p = K_{pt}W$ for the *prediction set*, where $W = [w^{(1)}, w^{(2)}, \cdots, w^{(m)}]$.

The inner products of the feature vectors between two diseases are hence denoted as $g(x, x') = u(x)^\top u(x')$. To represent all disease–disease similarities in the feature space, we defined the similarity matrix Q as $(Q)_{ij} = q(x_i, x_j) = u(x_i)^\top u(x_j)\,(1 \le i, j \le N)$. To split the matrix Q into several parts according to the *training set*, the *prediction set* and their interactions, we performed the following computations:

Training set versus *Training set:*

$$Q_{tt} = U_t U_t^\top = K_{tt} W W^\top K_{tt}^\top, \tag{4}$$

Prediction set versus *Training set:*

$$Q_{pt} = U_p U_t^\top = K_{pt} W W^\top K_{tt}^\top, \tag{5}$$

Prediction set versus *Prediction set:*

$$Q_{pp} = U_p U_p^\top = K_{pt} W W^\top K_{pt}^\top. \tag{6}$$

In these computations, we determined the $n \times m$ weight matrix W for which Q_{tt} fits G_{tt} as much as possible. If we set $H = WW^\top$, this problem can be addressed by finding the H that minimizes the difference between G_{tt} and Q_{tt}, thus avoiding considerable computational burdens for computing W itself, even if m is infinite. The $H(= WW^\top)$ minimizes

$$L = ||G_{tt} - K_{tt} H K_{tt}^\top||_F^2, \tag{7}$$

where $||\cdot||_F$ indicates the Frobenius norm. We can rewrite the above equation in the trace form as follows:

$$L = tr\{(G_{tt} - K_{tt} H K_{tt}^\top)(G_{tt} - K_{tt} H K_{tt}^\top)^\top\}. \tag{8}$$

Here we introduce regularization in KMR by finding $H(= WW^\top)$ that minimizes the following penalized loss function:

$$L = ||G_{tt} - K_{tt} H K_{tt}||_F^2 + \lambda PEN(H), \tag{9}$$

where λ is a regularization parameter and $PEN(H)$ is a penalty term for H and is defined as $PEN(H) = 2tr(HK_{tt})$. In this case, the optimization problem is reduced to finding H, which minimizes

$$L = tr\left\{(G_{tt} - K_{tt} H K_{tt}^\top)(G_{tt} - K_{tt} H K_{tt}^\top)^\top\right\} + 2\lambda tr\{HK_{tt}\}. \tag{10}$$

The derivative of L with respect to H is given by the following equation:

$$\frac{1}{2}\frac{\partial L}{\partial H} = -K_{tt} G_{tt} K_{tt} + K_{tt}^2 H K_{tt}^2 + \lambda K_{tt}$$

Therefore, the solution of the above penalized optimization problem is obtained as $H = {K_{tt}}^{-1}(G_{tt} - \lambda K_{tt}^{-1})K_{tt}^{-1}$. The penalty used is only justified for positive semidefinite matrices, which can be generated at least for sufficiently small λ values. Therefore, we compute the feature-based similarity matrix Q involving the prediction set as follows:

Prediction set versus *Training set:*

$$Q_{pt} = U_p U_t^\top = K_{pt} K_{tt}^{-1}(G_{tt} - \lambda K_{tt}^{-1}), \tag{11}$$

Prediction set versus *Prediction set:*

$$Q_{pp} = U_p U_p^\top = K_{pt} K_{tt}^{-1}(G_{tt} - \lambda K_{tt}^{-1}) K_{tt}^{-1} K_{pt}^\top. \tag{12}$$

Using Q_{pt} and Q_{pp}, we can predict the missing entries in the kernel matrix G that correspond with G_{pt} and G_{pp}.

We set the regularization parameter lambda by performing the fivefold cross-validation experiments, following the previous study (Yamanishi *et al.*, 2007). We evaluated the accuracy scores by varying the lambda parameter little by little, and obtained the highest performance when the lambda value was around 0.1. Thus, the lambda was set to 0.1 in this study.

3.3 Performance evaluation

We evaluated the performance by performing fivefold cross-validation experiments using receiver operating characteristic (ROC) curves that were derived from plots of true-positive rates versus false-positive rates, and precision–recall (PR) curves that were derived from plots of precision (positive predictive values) versus recall (sensitivity). We summarized performance using the area under ROC curve (AUC) scores in which 1 indicates perfect inference and 0.5 indicates purely random inference, and area under the PR curve (AUPR) scores for which 1 indicates perfect inference and the ratio of positive examples in the therapeutic target or drug-sharing data indicates random inference. We calculated the prediction scores for all disease–disease pairs, split them into ICD-11 disease categories, and evaluated the AUC scores for disease pairs within the same disease category or across disease categories.

To assess the efficacy of our network-based similarity, we prepared a profile-based similarity as a baseline. In the profile-based method, disease similarities were calculated by correlation coefficients based on disease-specific gene expression profiles. We transformed disease-specific gene expression profiles into binary profiles by distinguishing genes with high and low expression levels. Subsequently, we calculated Jaccard similarity scores. The profile-based method calculates disease similarities based on the overlap of disease-related genes between diseases. In contrast, the network-based method can calculate disease similarities even if there is no overlap of disease-related genes between diseases. There are many diseases that are characterized by molecular networks of genes rather than individual genes, as shown in the previous study (Menche *et al.*, 2015).

4. Results

4.1 Therapeutic target-sharing disease pairs have higher similarities

To test whether a disease is more likely to be proximal to other diseases that share therapeutic targets, we compared network similarities between disease pairs that did and did not share therapeutic targets.

Figure 1A shows distributions of network-based similarities between disease pairs, in which disease pairs are distinguished in terms of therapeutic target sharing. Among the possible 3081 disease pairs, 201 (6.5%) shared at least one therapeutic target. Disease pairs with shared therapeutic targets tended to have higher similarities than disease pairs that did not share gene targets in the mixture network ($P < 2.2e\text{-}16$; Kolmogorov-Smirnov test). Especially, a strong tendency was observed in 'Y2H interactions', 'protein complexes', 'kinase–substrate pairs', 'signal interactions' ($P < 2.2e\text{-}16$; Kolmogorov-Smirnov test). These findings suggest that disease pairs

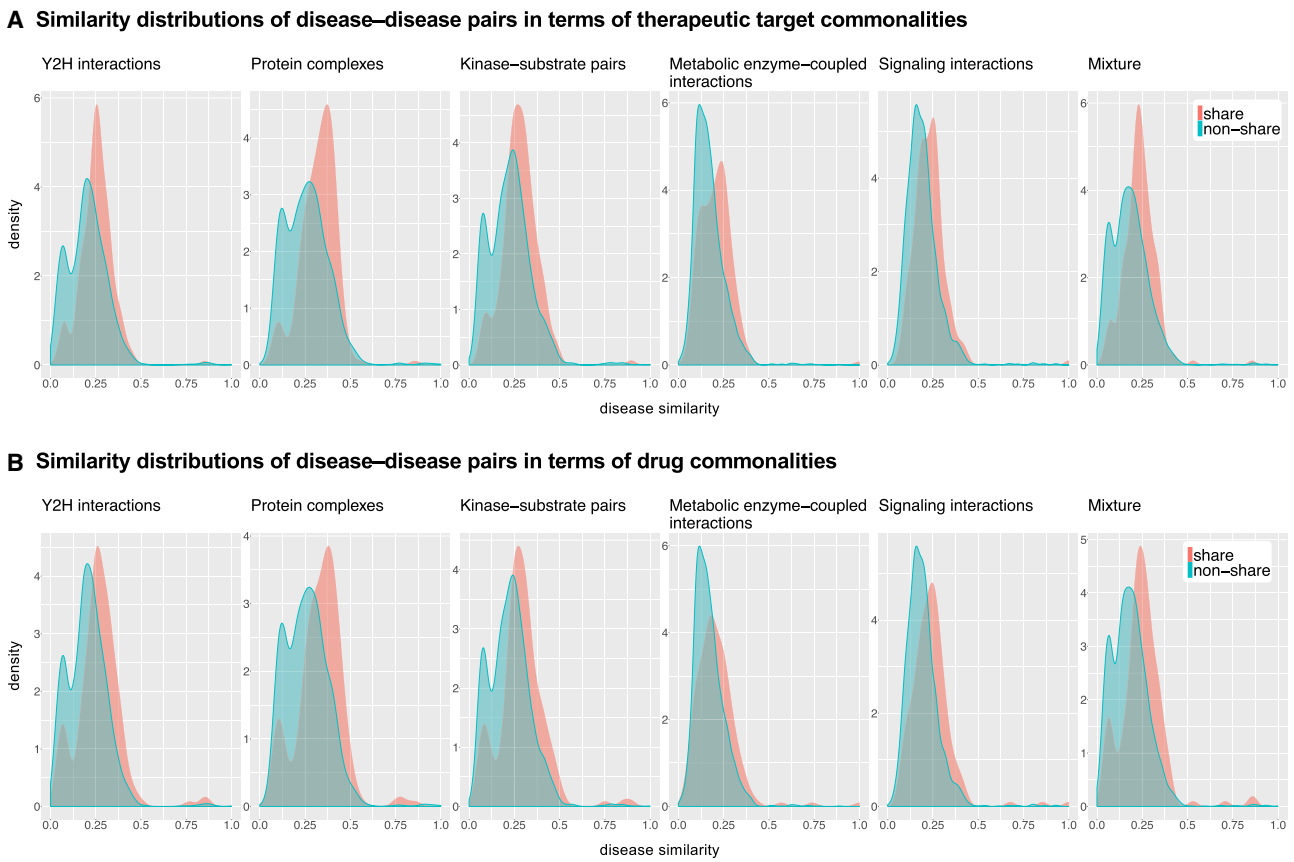

Fig. 1. Comparisons of network-based disease similarities between therapeutic target-sharing disease pairs and non-sharing pairs (**A**) and between drug-sharing disease pairs and non-sharing pairs (**B**). Distributions coloured in pink and blue correspond with disease pairs that share and do not share therapeutic targets or drugs, respectively. The horizontal axes indicate network-based disease similarities as calculated using 'Y2H interactions', 'protein complexes', 'kinase–substrate pairs', 'metabolic enzyme-coupled interactions', 'signal interactions' and all of those interactions (mixture)

with highly network-based similarities tend to share therapeutic targets.

4.2 Drug-sharing disease pairs have higher similarities

We determined whether diseases are more likely to be proximal to others when they share drugs. Figure 1B shows the distributions of network-based similarities for disease pairs that were distinguished in terms of sharing of approved drugs. In all possible 3081 disease pairs, 168 (5%) shared at least one drug. As for therapeutic target sharing (see Fig. 1A), drug-sharing disease pairs had higher similarities than those that did not share drugs ($P < 2.2e\text{-}16$; Kolmogorov-Smirnov test for all types of human interactomes). We show that the disease pairs that share approved drugs have higher topological similarities.

4.3 Performance evaluation of predictions of therapeutic target commonalities among diseases

Here we evaluated the performance for predicting therapeutic target commonalities among diseases. We compared the performance between unsupervised approach and supervised approach. As gold standard data, we used 201 disease pairs with shared therapeutic targets, including 37 diseases.

Table 2A shows AUC and AUPR scores for the unsupervised and supervised approaches with the baseline profile-based and network-based disease similarities. We used the Jaccard similarity score in the profile-based method, because the Jaccard similarity score tended to work better than other similarity scores such as Pearson and Spearman similarities. Supplementary Table S1 shows the results of the performance comparison between Pearson, Spearman and Jaccard similarities. The AUC and AUPR values are always higher in supervised algorithms than in their unsupervised counterparts. The profile-based and network-based similarities were comparable in terms of prediction of therapeutic target commonalities. We also evaluated the performance by precision at low recall, and provided the experimental results in Supplementary Table S2. The precision scores at low recall (i.e. 5 and 10%) with supervised learning were higher than those with unsupervised learning in most cases. In the case of supervised learning, the network-based methods achieved higher precision at low recall, compared with the profile-based method. For instance, the network-based methods with Y2H and signalling interaction networks achieved the highest precision at the 5% recall in the prediction tasks for both therapeutic targets and drugs. These results suggest the usefulness of the network-based method with supervised learning.

The prediction using the mixture interactome provides the best accuracy among network-based methods (AUC = 0.750), which is because the number of constructed interactions was highest for the mixture network. Metabolic enzyme-coupled interactions also had comparable AUC scores (AUC = 0.749) to those of mixture interactions, despite lower numbers of network proteins than in mixture networks. These results suggest that proteins involved in metabolic enzyme-coupled interactions tend to be therapeutic targets.

We further investigated whether disease categories can be easily predicted by specific networks. Each heatmap in Figure 2 corresponds to the performance evaluation with a kind of molecular interaction network. For example, the performance for ICD-11 category V diseases (endocrine, nutritional or metabolic diseases) was low in some cases (e.g. 'kinase–substrate pairs', 'protein complexes' and 'mixture' for the network-based method and the profile-based

Table 2. Performance evaluation for the prediction of commonalities of therapeutic targets and drugs

	Supervised learning		Unsupervised learning	
Method	AUC	AUPR	AUC	AUPR
(A) Therapeutic target commonality				
Profile-based method	**0.761**	**0.526**	0.685	0.408
Network-based method				
Mixture	0.750	**0.526**	0.692	0.423
Y2H interactions	0.715	0.510	0.682	0.417
Protein complexes	0.718	0.469	0.699	0.423
Kinase–substrate pairs	0.711	0.476	**0.700**	**0.438**
Metabolic enzyme-coupled interactions	0.749	0.508	0.660	0.422
Signalling interactions	0.718	0.482	0.674	0.421
(B) Drug commonality				
Profile-based method	0.836	0.087	0.690	0.032
Network-based method				
Mixture	**0.837**	0.099	**0.713**	0.039
Y2H interactions	0.779	0.088	0.693	0.037
Protein complexes	0.716	0.055	0.707	0.037
Kinase–substrate pairs	0.662	0.022	0.687	0.035
Metabolic enzyme-coupled interactions	0.807	0.076	0.685	**0.041**
Signalling interactions	0.819	**0.114**	0.711	0.039

Note: The best score for each method is highlighted in bold.

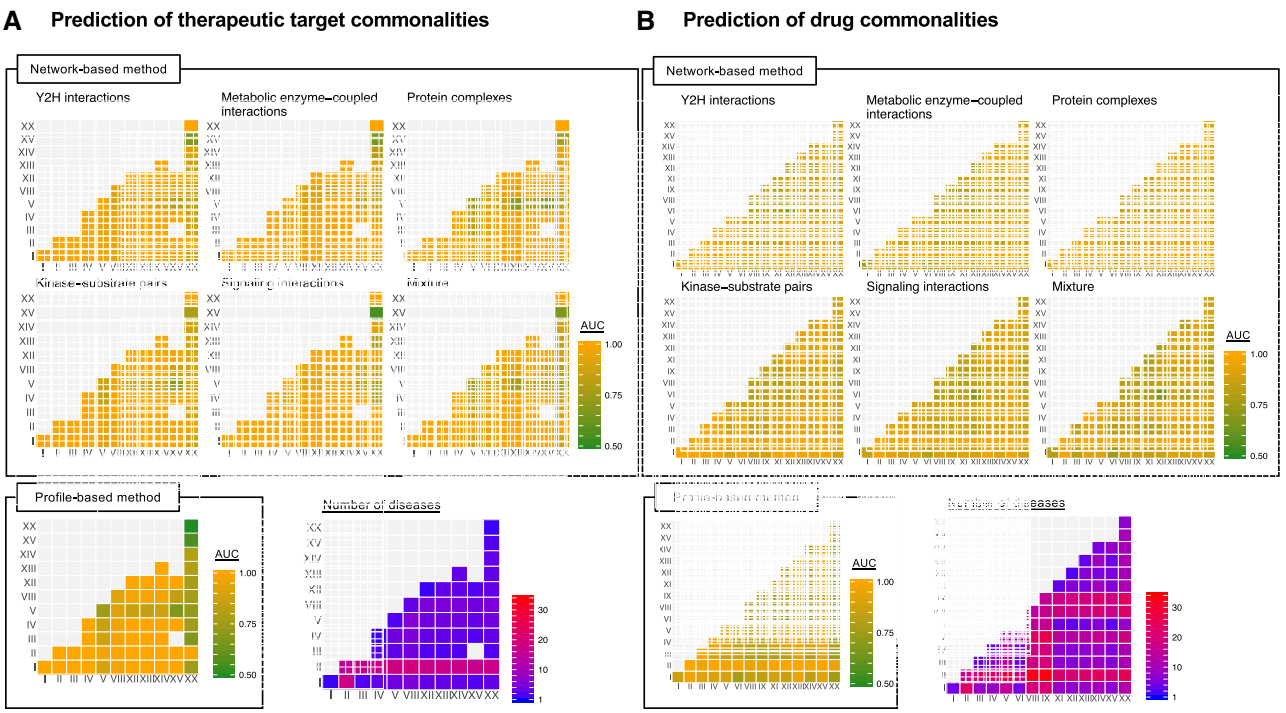

Fig.2. Performance evaluation based on ICD-11 disease chapters for therapeutic target commonality (**A**) and drug commonality (**B**). Roman numbers are showing the ICD-11 disease chapters. The green to yellow colour shows the AUC score for combination of ICD-11 diseases. The blue to red colour shows the number of diseases used for calculating the AUC. Chapter I: certain infectious or parasitic diseases; chapter II: neoplasms; chapter III: diseases of the blood or blood-forming organs; chapter IV: diseases of the immune system; chapter V: endocrine, nutritional or metabolic diseases; chapter VI: mental, behavioural or neurodevelopmental disorders; chapter VIII: diseases of the nervous system; IX: diseases of the visual system; chapter XII: diseases of the respiratory system; chapter XIII: diseases of the digestive system; chapter XIV: diseases of the skin; chapter XV: diseases of the musculoskeletal system or connective tissue: chapter XX: developmental anomalies

method). In contrast, the performance for ICD-11 category V diseases did not make a difference in the other cases (e.g. 'metabolic enzyme-coupled interactions' and 'signalling interactions'). Therefore, the diseases in ICD-11 category V could be well characterized by the metabolic-related and signalling-related interactome, not by the protein-binding and gene-expression interactome. These results suggest that the characterization of disease–disease relationships depends heavily on the kind of molecular interaction networks, and the use of the mixture interactome does not always work the best, which are important findings in this study.

4.4 Performance evaluation of predictions of drug commonalities among diseases

Next, we evaluated the performance for predicting drug commonalities among diseases. As gold standard data, we used 40 disease pairs with shared drugs between 34 diseases.

Table 2B shows AUC and AUPR scores for the unsupervised and supervised approaches with the baseline profile-based and network-based disease similarities. The AUC score was the highest for the comprehensive human interactome (AUC = 0.837), which had a comparable AUC value and a higher AUPR value than those from baseline calculations. This result suggests that the interactome offers considerable predictive power for drug commonalities.

Figure 2B compares AUC scores between ICD-11 disease chapters. The AUC scores between ICD-11 chapter VIII (diseases of the nervous system) and chapter XI (diseases of the circulatory system) from complexes was higher than the others. Similarly, the pair of ICD-11 chapter XII (diseases of the respiratory system) and ICD-11 chapter XIV (diseases of the skin) gave slightly higher AUC values in 'protein complexes' and 'kinase–substrate pairs' than the other chapters, suggesting that some disease categories tend to be predicted by specific protein–protein networks.

4.5 Novel prediction of therapeutic target-sharing disease pairs

We predicted novel therapeutic target-sharing disease pairs by the profile-based and network-based methods. We used 37 diseases with information on therapeutic targets as a training set, where 201 pairs of diseases were known to share at least one therapeutic target. Overall, 43 novel disease pairs shared therapeutic targets (Supplementary Fig. S1 and Table S3). Our network-based prediction demonstrated more therapeutic target-sharing disease pairs than the profile-based method.

For example, the association of Parkinson's disease (H00057) and Lewy body dementia (LBD) (H00066) was predicted from both profile-based and network-based computations (Supplementary Table S3). This result suggests that these diseases have many overlapping genes and high similarities in the human interactome. In contrast, the association between adrenoleukodystrophy (H00176), involving ICD-11 chapter V (endocrine, nutritional or metabolic diseases) and idiopathic pulmonary fibrosis (IPF; H01299), involving ICD-11 chapter XII (diseases of the respiratory system), was only predicted in the Y2H network. Perhaps suitable combinations of network and disease pairs will be optimally predictive of therapeutic targets.

4.6 Novel prediction of drug-sharing disease pairs

Finally, we predicted novel drug-sharing disease pairs for 79 diseases. Among these, 40 were known drug-sharing associations between 34 diseases and were used as training data. Overall, we found novel 68 disease pairs with shared drugs (Fig. 3 red edges and Supplementary Table S4).

Although diseases of the same ICD-11 disease groups tended to share drugs (Fig. 3 grey edges), our network-based method predicted associations between different categories of ICD-11 diseases. For example, IPF (H01299), which is a disease of the respiratory system in ICD11 and is categorized in ICD-11 chapter XII (diseases of the respiratory system), may share drugs with atopic dermatitis (AD; H01358), which is categorized in ICD-11 chapter XIV (diseases of the skin). Indeed, prednisolone sodium phosphate (D00981), which was not in the KEGG Release 84.1 (released on December 1, 2017) from which we derived information concerning drugs for diseases, is approved for the treatment of both diseases in the KEGG database Release 91.0 (released on July 1, 2019). However, this association was only predicted by 'kinase–substrate pairs', indicating that these diseases only share kinase–substrate interactions.

The glucocorticoid receptor (GR) is a therapeutic target of prednisolone sodium phosphate and is located close to IPF related genes (green circles in Supplementary Fig. S2) in the kinases network. Although the gene encoding GR is not related to IPF in our dataset, a recent study demonstrated that GR expression was significantly lowered in lung tissues from IPF and might be a biomarker for IPF (Bin *et al.*, 2019). Similarly, the dopamine receptor D2 (DRD2), which is a therapeutic target of AD, was localized in neighbourhoods of the IPF module. It has been shown that inhibition of the dopamine receptor D1 (DRD1), which is a subtype of dopamine receptors, reverses fibrosis through YAP/TAZ signalling in mice. Hence, some drugs may be effective for both IPF and AD.

5 Discussion and conclusions

In this article, we present a novel method for revealing disease–disease relationships based on molecular interaction networks using machine learning. The originality of the approach lies in the quantitation of disease similarities from characteristics of five molecular interaction networks and disease-specific omics profiles. This approach revealed novel disease–disease relationships that share therapeutic targets and drugs. In particular, the KMR identified novel disease–disease associations, with greater effect than the unsupervised method. Our approach is expected to be useful for understanding disease–disease relationships. The information on predicted disease commonality pairs can be used to find new candidates for drugs and therapeutic targets. For example, if one disease in a predicted disease commonality pair has a known drug, the other disease (without drug information) in the disease commonality pair is predicted to have the same drug. Likewise, if one disease in a predicted disease commonality pair has a known therapeutic target, the other disease (without target information) in the disease commonality pair is predicted to have the same therapeutic target.

In this study, we demonstrated that topological molecular network characteristics of diseases that share therapeutic targets or drugs differ from those of diseases that do not share drugs or therapeutic targets. Considering disease similarities based on types of molecular networks helps to predict disease relationships. For example, the association between IPF and AD was predicted only in analyses of 'kinase–substrate pairs' (Fig. 3). In reality, these diseases share this drug and were added to the KEGG Release 91.0 (released on July 1, 2019) but were not present in KEGG Release 84.1 (released on December 1, 2017), from which we extracted information regarding drugs for diseases.

In recent pharmaceutical studies, network-based approaches have provided promising frameworks and novel insights that may accelerate drug discovery (Barabási *et al.*, 2004) by quantifying disease–disease (Menche *et al.*, 2015), drug–disease (Cheng *et al.*, 2018; Guney *et al.*, 2016) and drug–drug–disease relationships (Cheng *et al.*, 2019). More careful examinations are required, however, to decipher novel disease–disease associations that share therapeutic targets and drugs. Previously developed methods considered only disease-causing genes, and excluded disease-specific gene profiles based on mRNA expression levels. Because disruptions of coordinated gene expression will likely be more definitive of disease pathogenesis, and because disease-causing genes with mutations are rarely considered as therapeutic targets, further studies are required to consider gene expression profiles in predictions of therapeutic targets and drugs. In addition, the previous

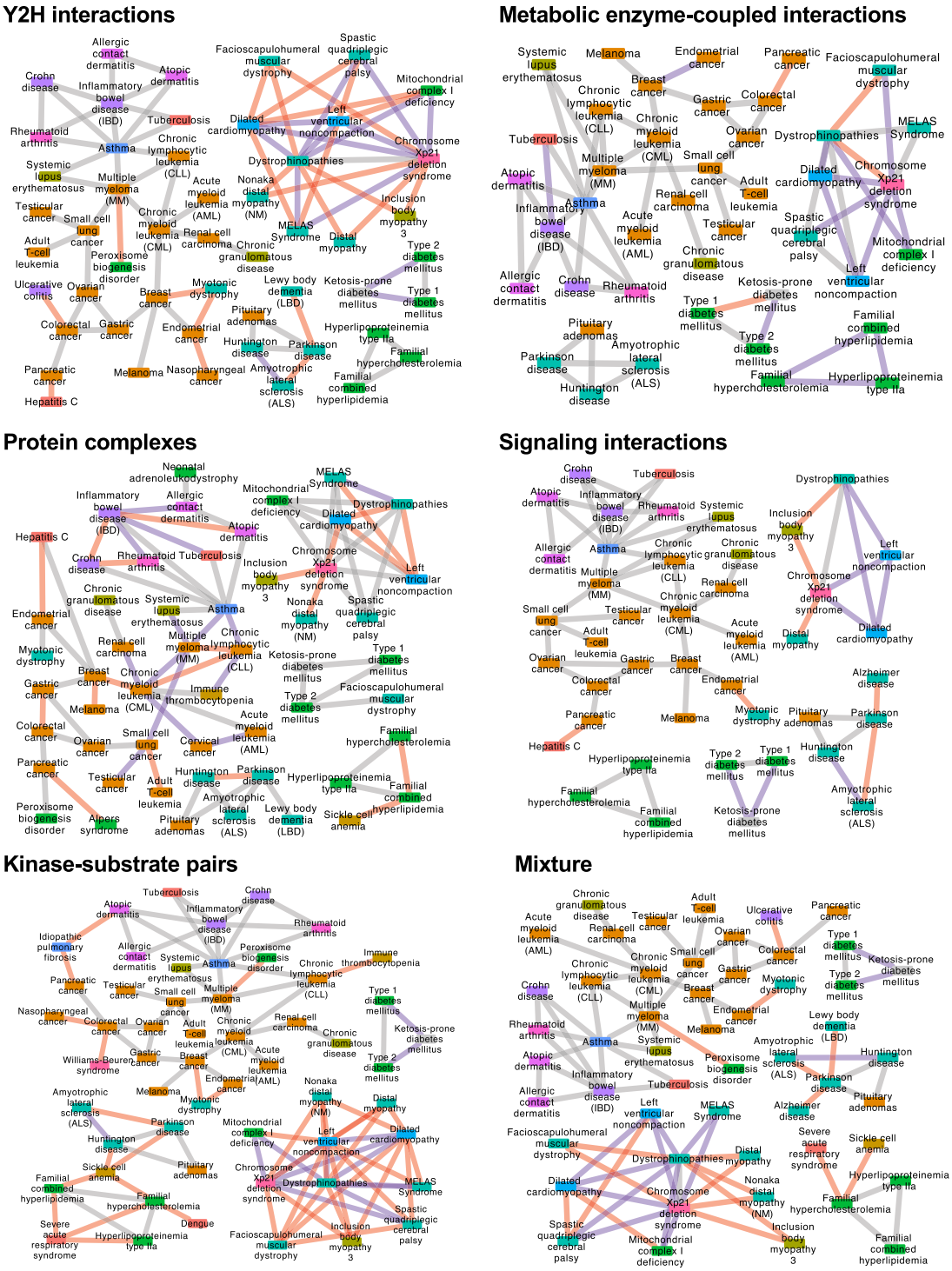

Fig. 3. Predicted associations of drug-sharing diseases. Gray, red, blue and purple edges indicate known associations, predicted by network-based method only, predicted by profile-based method only, and predicted by both network-based and profile-based methods, respectively. Details of ICD-11 chapters are described in Figure 2 legend

studies neglected to be attentive to types of molecular interactions in networks. Perhaps suitable types of networks will be predictive of new disease associations that share therapeutic targets and drugs, warranting studies of individual types of molecular interactions.

Funding

This work was supported by JST AIP-PRISM [grant number JPMJCR18Y5], Japan.

Conflict of Interest: none declared.

References

Amberger,J.S. and Hamosh,A. (2017) Searching Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic phenotypes. *Curr. Protoc. Bioinf.*, **58**, 1.2.1–1.2.12.

Barabási,A.-L. *et al.* (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Bin,Y.F. *et al.* (2019) Expression of GR-α and HDAC2 in steroid-sensitive and steroid-insensitive interstitial lung disease. *Biomed. Pharmacother.*, **118**, 109380.

Boolell,M. *et al.* (1996) Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *Int. J. Impot. Res.*, **8**, 47–52.

Cheng,F. *et al.* (2018) Network-based approach to prediction and population-based validation of *in silico* drug repurposing. *Nat. Commun.*, **9**, 2691.

Cheng,F. *et al.* (2019) Network-based prediction of drug combinations. *Nat. Commun.*, **10**, 1197.

Clark,N.R. *et al.* (2014) The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics*, **15**, 79.

Fazekas,D. *et al.* (2013) SignaLink 2—a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.*, **7**, 7.

Giurgiu,M. *et al.* (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.*, **47**, D559–D563.

Guney,E. *et al.* (2016) Network-based in silico drug efficacy screening. *Nat. Commun.*, **7**, 10331.

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

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Iwata,M. *et al.* (2019) Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm. *Bioinformatics*, **35**, i191–i199.

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

Kibbe,W.A. *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.

Mottaz,A. *et al.* (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics*, **9** Suppl 5, S3.

Menche,J. *et al.* (2015) Disease networks. Uncovering disease–disease relationships through the incomplete interactome. *Science*, **347**, 1257601–1257601.

Papadakis,M.A. *et al.* (2013) *Current Medical Diagnosis and Treatment 2014*, 53rd edn. McGraw-Hill Medical, San Francisco.

Rolland,T. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.

Ramos,E.M. *et al.* (2014) Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.*, **22**, 144–147.

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

Shimizu,Y. *et al.* (2008) Generalized reaction patterns for prediction of unknown enzymatic reactions. *Genome Inform.*, **20**, 149–158.

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

Venkatesan,K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.

Wang,Z. *et al.* (2016) Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nat. Commun.*, **7**, 12846.

Weimann,J. *et al.* (2000) Sildenafil is a pulmonary vasodilator in awake lambs with acute pulmonary hypertension. *Anesthesiology*, **92**, 1702–1712.

Yamanishi,Y. *et al.* (2007) Kernel matrix regression. In: *Proceedings of the 12th International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2007)*. hal-00133355.

Yang,X. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**, 805–817.

Yıldırım,M.A. *et al.* (2007) Drug–target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Yu,H. *et al.* (2011) Next-generation sequencing to generate interactome datasets. *Nat. Methods*, **8**, 478–480.