SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
DATA INTEGRATION
MACHINE LEARNING
PREDICTIVE MEDICINE

Received
10 September 2013

Accepted
23 October 2013

Published
15 November 2013

Correspondence and requests for materials should be addressed to B.Z. (blaz.zupan@fri.uni-lj.si) or N.P. (natasha@imperial.ac.uk)

Discovering disease-disease associations by fusing systems-level molecular data

Marinka Žitnik[1], Vuk Janjić[2], Chris Larminie[3], Blaž Zupan[1,4] & Nataša Pržulj[2]

[1]Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1000, Slovenia, [2]Department of Computing, Imperial College London, London, SW7 2AZ, United Kingdom, [3]Computational Biology, GlaxoSmithKline, Stevenage, Hertfordshire, SG1 2NY, United Kingdom, [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.

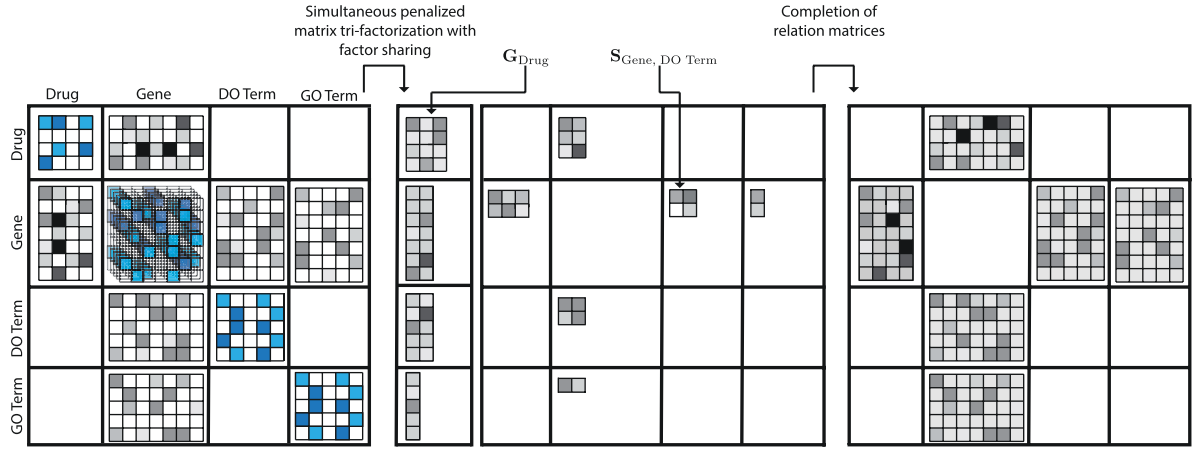
The advent of genome-scale genetic and genomic studies allows new insight into disease classification. Recently, a shift was made from linking diseases simply based on their shared genes towards systems-level integration of molecular data. Here, we aim to find relationships between diseases based on evidence from fusing all available molecular interaction and ontology data. We propose a multi-level hierarchy of disease classes that significantly overlaps with existing disease classification. In it, we find 14 disease-disease associations currently not present in Disease Ontology and provide evidence for their relationships through comorbidity data and literature curation. Interestingly, even though the number of known human genetic interactions is currently very small, we find they are the most important predictor of a link between diseases. Finally, we show that omission of any one of the included data sources reduces prediction quality, further highlighting the importance in the paradigm shift towards systems-level data fusion.

Disease Ontology (DO)[1] is a well established classification and ontology of human diseases. It integrates disease nomenclature through inclusion and cross mapping of disease-specific terms and identifiers from Medical Subject Headings (MeSH)[2], World Health Organization (WHO) International Classification of Diseases (ICD)[3], Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[4], National Cancer Institute (NCI) thesaurus[5] and Online Mendelian Inheritance in Man (OMIM)[6]. It relates and classifies human diseases based on pathological analysis and clinical symptoms. However, the growing number of heterogeneous genomic, proteomic, transcriptomic and metabolic data currently does not contribute to this classification. Understanding of even the most straightforward monogenic classic Mendelian disorders is limited without considering interactions between mutations and biochemical and physiological characteristics. Hence, redefining human disease classification to include evidence from heterogeneous data is expected to improve prognosis and response to therapy[7]. In this paper we examine whether inclusion of modern molecular level data can improve disease classification.

Several studies have reported on efforts and benefits of relating human diseases through their molecular causes. Loscalzo et al.[7] catalogued diseases through a network-based analysis of associations among genes, proteins, metabolites, intermediate phenotype and environmental factors that influence pathophenotype. Gulbahce et al.[8] constructed a "viral disease network" of disease associations to decipher the interplay between viruses and disease phenotypes. They uncovered several diseases that have not previously been associated with infection by the corresponding viruses. A similar approach was used by Lee et al.[9] to gain insights into disease relationships through a network derived from metabolic data instead of virological implications. They demonstrated that known metabolic coupling between enzyme-associated diseases reveal comorbidity patterns between diseases in patients. Goh et al.[10] studied the position of disease genes within the human interactome in order to predict new cancer-related genes. Conversely, a gene-centric approach to disease association discovery was used by Linghu et al.[11]: they took 110 diseases for which a set of disease genes are known, and compared gene sets and their positions within the gene network to infer associations of related diseases. More details can be found in two recent surveys of current network analysis methods aimed at giving insights into human disease[12,13], as well as in a review of different data sources that can provide complementary disease-relevant information[14].

A challenge in relating diseases and molecular data is in the multitude of information sources. Disease profiling may include data from genetics, genomics, transcriptomics, metabolomics or any other omics, all potentially

A Graphical representation of data fusion



B Algorithm for disease association prediction and class assignment

- **Input:** A sequence of matrix factors from 15 repetitions of factorisation $\mathbf{G}_2^{(i)}$ for $1 \leq i \leq 15$.
- **Output:** Consensus matrix, $\bar{\mathbf{C}}$, and a set, $\mathcal{D}$, of disease classes, D.

1. Repeat the following for each matrix factor $\mathbf{G}_2^{(i)}$ for $1 \leq i \leq 15$.
   (a) For each disease j compute its class as $\arg\max_m \mathbf{G}_2^{(i)}(j, m)$.
   (b) Compute connectivity matrix $\mathbf{C}^{(i)}$ from class assignment such that $\mathbf{C}^{(i)}(r, s)$ is set to 1 if disease r and s were assigned the same class in step a.
2. Compute consensus matrix as $\bar{\mathbf{C}} = \frac{1}{15}\sum_i \mathbf{C}^{(i)}$.
3. Extract new disease classes, $\mathcal{D} = \{D \mid D \subset \{\text{Disease Ontology IDs}\} \wedge \forall i, j \in D \wedge i \neq j : \bar{\mathbf{C}}(i, j) = 1\}$.

Figure 1 | Data fusion. Panel **A** is a graphical representation of our data fusion by matrix factorisation approach to discovering disease-disease associations. The shown block-based matrix representation exactly corresponds to the data fusion schema in Figure 3-A. We combine 11 data sources on four different types of objects (see Methods): drugs, genes, Disease Ontology (DO) terms and Gene Ontology (GO) terms. These data are encoded in two types of matrices: constraint matrices, which relate objects of the same type (such as drugs if they have common adverse effects) and are placed on the main diagonal (illustrated by matrices with blue entries); and relation matrices, which relate objects of different types and are placed off the main diagonal (illustrated by matrices with grey entries). Our data fusion approach involves three main steps. First, we construct a block-based matrix representation of all data sources used in our study (panel A, left). The molecular data encoded in these matrices are sparse, incomplete and noisy (depicted by different shades of blue and grey) and some matrices are completely missing because associated data sources are not available (e.g. no link between GO terms and drugs). In the second step, we simultaneously decompose all relation matrices as products of low-rank matrix factors and use constraint matrices to regularise low-rank approximations of relation matrices. The key idea of our data fusion approach is sharing low-rank matrix factors between relation matrices that describe objects of common type. The resulting factorised system (panel A, middle) contains matrix factors that are specific to every type of objects (four matrices in left part; e.g. $\mathbf{G}_{\text{Drug}}$), and matrix factors that are specific to every data source (six matrix factors in right part; e.g. $\mathbf{S}_{\text{Gene, DO Term}}$). Thus, low-rank matrix factors capture source- and object type-specific patterns. Finally, we use matrix factors to reconstruct relation matrices and complete their unobserved entries (panel A, right). Panel **B** shows the algorithm for assigning diseases to classes and obtaining disease-disease association predictions.

related to susceptibility, progress and manifestation of disease. Such data may be related on their own: for example, information on transcription factor binding sites, gene and protein interactions, drug-target associations, various ontologies and other less-structured knowledge bases, such as literature repositories, are all inter-dependent and it is not trivial to integrate them in a way that will yield new information about diseases. This stresses the need for an integrated approach of current models to exploit all these heterogeneous data simultaneously when inferring new associations between diseases[13].

Data from heterogeneous sources of information can be integrated by *data fusion*[15]. Common fusion approaches follow early or late integration strategies, combining inputs[16] or predictions[17], respectively. Another and often preferred approach is an intermediate integration, which preserves the structure of the data while inferring a single model[18–20]. An excellent example of intermediate integration is multiple kernel learning that convexly combines several kernel matrices constructed from available data sources[15,21]. Data fusion has been successfully applied for tasks such as gene prioritisation[15,21,22], or gene network reconstruction and function prediction[16,23]. To our knowledge, we present the first application of data fusion to disease association mining.

We choose the intermediate data fusion approach for its accuracy of inferring prediction models (i.e. how well a model can learn to predict disease-disease associations) and the ability to explicitly measure the contribution of each data set to the extracted knowledge[18,19]. Kernel-based fusion can only use data sources expressed in the "disease space", i.e. all data sources have to be expressed as kernel matrices encoding relationships between diseases, which may incur loss

of information when transforming circumstantial data sources into appropriate feature space. In our study, most of the data sources are only indirectly related to diseases, hence we employ an alternative and recently proposed intermediate data fusion algorithm by matrix factorisation[24], which has an accuracy comparable to kernel-based fusion approaches, but can treat all data sources directly (i.e. no transformation of data into "disease space" is necessary). The key idea of our data fusion approach lies in sharing of low-rank matrix factors between data sources that describe biological data of the same type. For instance, genes are one data type which can be linked to other data types such as Gene Ontology (GO) terms or diseases through two distinct data sources, namely GO annotations and disease-gene mapping. The fused factorised system contains matrix factors that are specific to every molecular data type, as well as matrix factors that are specific to every data source. Thus, low-rank matrix factors can simultaneously capture both source- and object type-specific patterns.

We report on the ability of our recently developed data fusion approach to mine human disease-disease associations. Starting from Disease Ontology, we revise the links between diseases using related systems-level data, including protein-protein and genetic interactions, gene co-expressions, metabolic data, drug-target relations, and other (see Methods). By fusing these data we identify several disease-disease associations that were not present in Disease Ontology and validate their existence by finding strong support in the literature and significant comorbidity effects in associated diseases. We also quantify the contribution of each molecular data source to the integrated disease-disease association model.

Results

We fuse systems-level molecular data by using our recently developed matrix-factorisation approach (described in Methods) to gain new insight into the current state-of-the-art human disease classification. This large-scale data integration results in 108 highly reliable disease classes (each corresponding to a clique in the consensus matrix, $\bar{\mathbf{C}}$; see Methods section and Algorithm in Figure 1-B). Size distribution of the 108 disease classes is as follows: 60 disease classes contain 2 diseases; 31 disease classes contain 3 or 4 diseases; 9 disease classes contain 5, 6 or 7 diseases; 5 disease classes contain 8, 9 or 10 diseases; 2 disease classes contain 11 or 17 diseases; and 1 disease class contains 146 diseases. For each class we examine the associations between its member diseases to inspect how the obtained classes align with currently accepted disease classification.

Using Disease Ontology (DO) and literature curation, we find that the 107 smaller classes successfully capture closely-related diseases that are also placed near each other in DO (see below for details). Also, we find that in the largest identified disease class (i.e. the one containing 146 diseases), the most represented major disease is cancer (31.5%), followed by nervous system diseases (14.4%), inherited metabolic disorders (9.6%) and immune system diseases (5.5%). This class primarily contains diseases of anatomical entity (45.2%), cellular proliferation (25.4%) and metabolic diseases (14.3%), with other major concepts of DO being rarely represented. The large size of this class may reflect the following underlying biases in various data sources – its constituents represent either larger majority groups in DO, or minority groups at a lower level of ontology:

- diseases of anatomical entity, because diseases are often described based on tissue/organ;
- cellular proliferation, because of the heavy enrichment of cancers and the sub-classification of these into many variant diseases, also possibly driven by rich gene/pathway annotation around cell cycle and proliferation;
- metabolic diseases, because of significant representation of metabolic diseases and significant understanding of metabolic pathways. Metabolic disease is a primary focus for systems modelling and simulation, as much is known from pathways and a wealth of omics data available.

Since the obtained distribution appears unbalanced due to one large class containing 146 diseases, we further decompose that class by repeating data fusion analysis on its disease members. This effectively gives us a multi-layer hierarchical breakdown of disease classes (see Figure 2). The large class is broken down into 10 classes (only those observed in all 15 inferred models are taken into account; see Methods section). The distribution of disease class sizes is: 9 disease classes with 2 or 3 diseases, and 1 disease class with 51 diseases. The diseases captured by the 9 smaller classes are: two classes consist of cancer diseases, three consist of inherited metabolic disorders, one contains nervous system diseases, two contain respiratory system diseases, and the last one has cardiovascular system diseases. The largest disease class (containing 51 disease members) is further decomposed into 8 disease classes. The distribution of disease class sizes at this level of hierarchy is: 7 disease classes with 2 or 3 diseases, and 1 disease class with 18 diseases. The diseases captured by the 7 smaller classes are: two classes with immune system diseases, one class with cognitive disorders, one class with acquired metabolic diseases, one with cancer, and the last three were split between cognitive disorders and metabolic diseases. The largest class (containing 18 disease members; again, under the most stringent agreement threshold; see Methods) is finally decomposed into six conserved diseases (the remaining 12 diseases grouped less reliably under our stringent threshold): lung metastasis, dysgerminoma, serous cystadenoma (cellular proliferation and cancer), abetalipoproteinemia (metabolic disorder), related factor XIII deficiency and plasmodium falciparum malaria.

Diseases in captured classes exhibit significant comorbidity. A comorbidity relationship exists between diseases whenever they affect the same individual substantially more than expected by chance. We want to know whether diseases assigned to the same disease class by our data fusion method exhibit higher comorbidity than diseases assigned to different classes. Hidalgo et al.[25] proposed two comorbidity measures (http://barabasilab.neu.edu/projects/hudine) to quantify the distance between two diseases: a relative risk (defined below) and Pearson's correlation between prevalences of two diseases (ϕ). A *relative risk* (RR) of two diseases is defined as the fraction between the number of patients diagnosed with both diseases and random expectation based on disease prevalence. Expressing the strength of comorbidity is difficult because different statistical distance measures are biased to under- or over-estimating the relationships between rare and prevalent diseases. The RR overestimates associations between rare diseases and underestimates associations involving highly prevalent diseases, whereas ϕ has low values for diseases with extremely different prevalence, but is good at recognising comorbidities between disease pairs of similar prevalence.

We find that 66 (out of 107) disease classes have a significantly higher comorbidity than what would be expected by chance (p-value < 0.001 with Bonferroni multiple comparison correction applied to all p-values). We assess the statistical significance by randomly sampling disease sets of the same size as the disease class in question, and computing the comorbidity enrichment scores of the sampled sets according to the two comorbidity measures, RR and ϕ, as proposed by Hidalgo et al.[5]. The enrichment score is then computed as the mean of comorbidity values between all disease pairs in a disease class. For subsequent layers of hierarchical decomposition of the largest disease class (i.e. the one containing 146 diseases), we find that: 7 out of 10 first level disease classes have a significantly higher comorbidity (measured by both RR and ϕ) than what would be expected by chance; comorbidity data was available for only 3 out of 8 second-level disease classes, and 2 of them exhibited significantly higher comorbidity than what would be expected by chance.

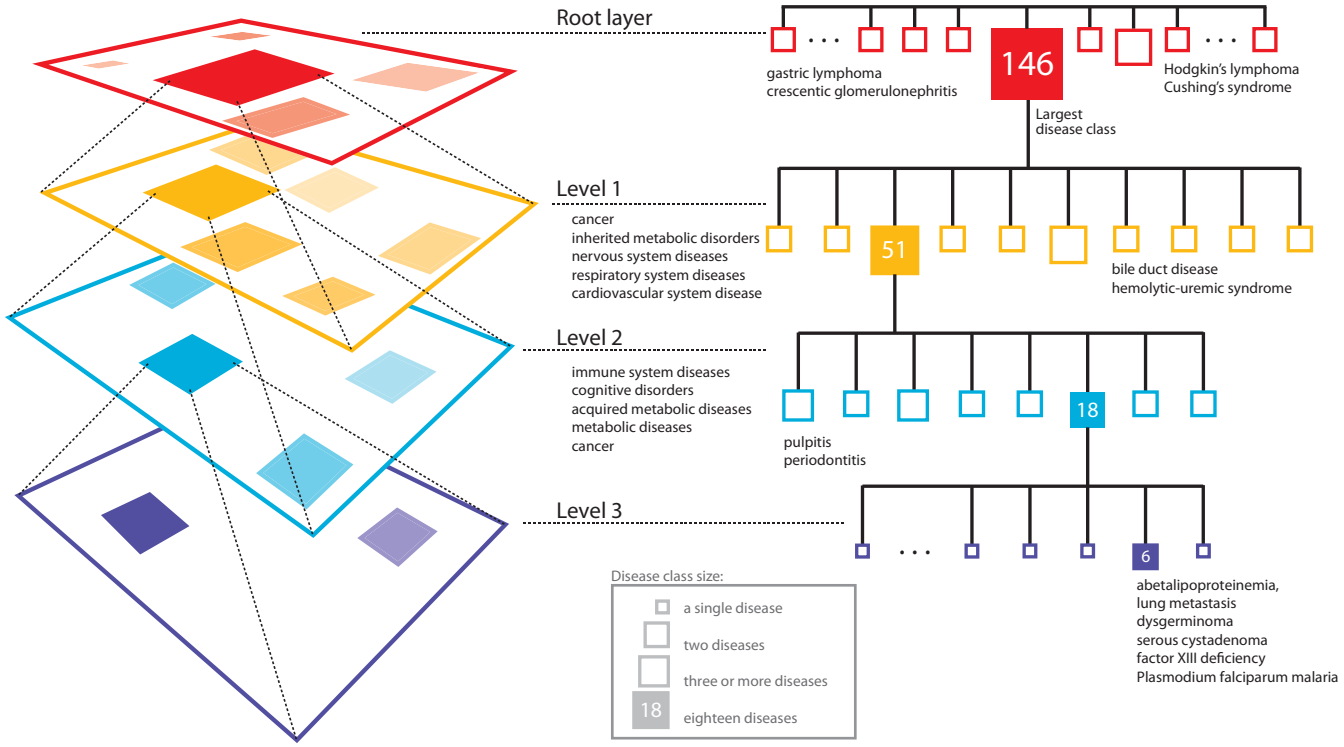


Figure 2 | Multi-layered hierarchical decomposition of disease classes. Our analysis yields 108 disease classes using the most stringent threshold for predicting disease-disease associations. Identified classes are rather small and each class contains at most 17 diseases with the exception of the largest disease class that consists of 146 diseases (at root layer). We further decompose the largest class by re-running the data fusion process on set of diseases that are in the largest class in order to identify its fine-grained structure (level one). We repeat data fusion analysis using this top-down strategy two more times (levels two and three), which results in a hierarchical decomposition of most reliable disease classes (see Methods).

Evaluating disease classes through Disease Ontology. To see how well our fusion approach captures disease-disease associations already present in the semantic structure of DO, we look at the overlap between 107 disease classes (again, we perform enrichment analysis of the largest above-described class separately, see below) and find that 79 classes have at least 80% of disease members directly connected in DO via *is_a* relationship; an example of one such disease class is given in Figure 3-B. We assess the statistical significance of such a high number of classes being enriched in known relations from DO by computing the *p*-value as follows. First, we remove all DO-related information (i.e. we remove the constraint matrix $\mathbf{\Theta}_2$; see Methods) and then we perform the data fusion again without any prior information on relationships between diseases. We find that such a high number of classes is unlikely to be enriched in known relations from DO by chance (p-value < 0.001).

This result is very interesting as it indicates that DO could, in principle, be reconstructed from molecular data only. Our findings suggest that disease classification derived from pathological analysis and clinical symptoms (DO) can be largely reproduced by considering *only* molecular data. In other words, data fusion of different types of evidence could be used to infer a hierarchy of disease relations whose coverage and power might be very similar to those of the manually curated DO.

The decomposition of the largest disease class yields similar results: 5 out of 9 first-level classes have their members directly linked in DO via *is_a* relationships; 4 out of 7 second-level disease classes have their members directly linked in DO via *is_a* relationships; the third-level class of size six does not significantly overlap with the DO graph, but is partially supported by literature[26].

Finding new links between diseases. In addition to examining classes of multiple diseases, we can use our fused model to rank individual disease-disease associations based on supporting molecular evidence, and make novel predictions linking previously seemingly unrelated diseases. Among all the highest-ranked disease-disease associations in the fused model (i.e. disease pairs from the most stable classes – obtained in step 3 of Algorithm in Figure 1-B – with less than 6 disease members), we find 14 associations not recorded in Disease Ontology. We perform literature curation and find evidence for *all* 14 of the predicted disease associations (Table 2). Such high accuracy is due to our choice to take a highly stringent approach that requests the association to be observed in all 15 of the inferred models (see Methods for details). Comorbidity data were available for 4 out of 14 predicted disease associations and all 4 of these disease-disease associations were found to have significantly high comorbidity: (DOID:11198, DOID:12336), (DOID:12252, DOID:8543), (DOID:423, DOID:13166), and (DOID:11202, DOID:11335).

Contribution of each data source to the fused model. We have seen that data fusion can successfully retrieve existing and uncover new associations between diseases. Now we examine the contribution of each individual data source to the final disease-disease association model. We estimate the relative importance of each of the fused data sources in predicting disease associations by comparing the quality of the inferred model that includes the data source, to the quality of the model that excludes it. The measured quality is represented by a tuple of residual sum of squares (RSS; lower values are better) and explained variance (Evar; higher values are better; see[24] for details) of gene-disease relationship matrix $\mathbf{R}_{12}$ (see Methods). So an increase in RSS and a decrease in Evar hinder the quality of the inferred model, and conversely, a decrease in RSS and an increase in Evar improve the quality of the inferred model. We find that omission of each of the five data sources that specify interactions between genes ($\mathbf{\Theta}_1^{(1)}, \ldots, \mathbf{\Theta}_1^{(5)}$) reduces the overall quality of the model. Surprisingly, the largest model degradation is observed in the absence of genetic interactions when Evar drops by 9.5% and RSS increases by

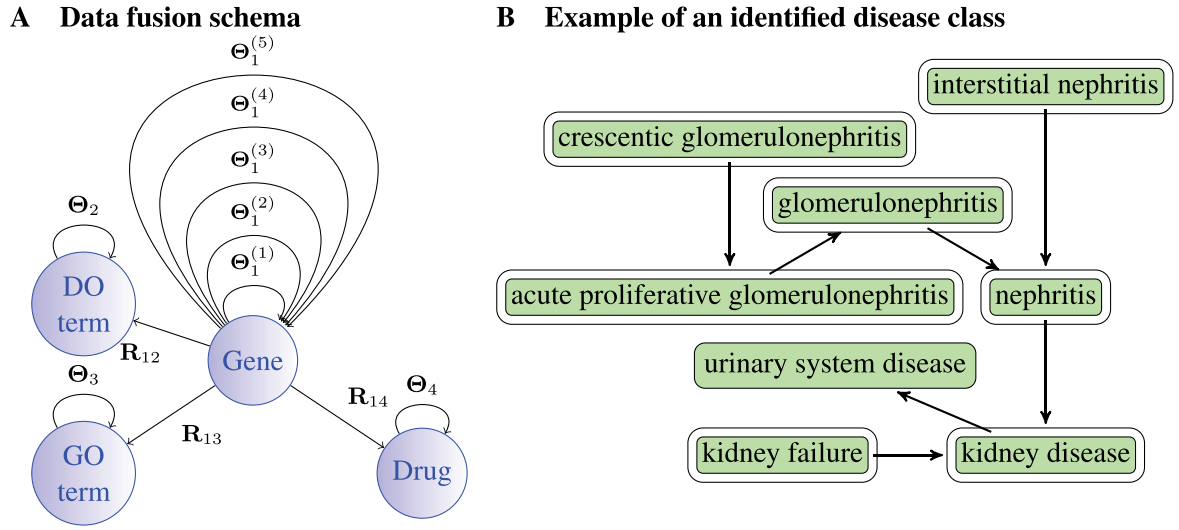


Figure 3 | System-level data fusion approach to disease re-classification. Panel **A** shows the relationships between data sources: nodes represent four types of objects, i.e. genes, GO terms, DO terms and drugs; arcs denote data sources that relate objects of different types (relation matrices, $\mathbf{R}_{ij}$, $i \neq j$), or objects of the same type (constraints, $\mathbf{\Theta}_i$). Panel **B** shows a disease class predicted by data fusion overlaid with a DO graph. Members of the disease class are outlined. This illustrates the ability of data fusion to successfully capture real disease classes: diseases associated with crescentic glomerulonephritis are presented.

13.3%. This result is unexpected, because the number of available genetic interactions is small (511). This may confirm the proposed importance of genetic interactions and functional buffering as being critical for understanding disease evolution and for design of new therapeutic approaches[27]. Although the dataset of genetic interactions is currently small, the observed interactions are more likely to be causative as opposed to correlative and may therefore have less noise associated, hence they appear to be more informative and have a larger importance on relationships between diseases than other data sources. Exclusion of other sources results in a smaller decrease in quality (Table 3), but nevertheless, these results confirm that all of the fused data sources contribute to the quality of the model.

Discussion

We integrate a wide range of modern systems-level molecular interaction and ontology data using our recently proposed data-fusion approach, and apply it to finding relationships between diseases previously unrecorded in DO. We validate our findings through comorbidity data and literature curation to demonstrate that such a systems-level integration can recover known and successfully identify currently unrecorded relationships between diseases.

When searching for disease-disease associations not present in DO, we considered only those associations that are present in all of the inferred models. This conservative approach gave us 14 disease-disease association predictions which we validated through literature and comorbidity data. Relaxing the threshold of association to be predicted, i.e. requiring a disease-disease association to be present in 95%, 90%, 85% or fewer of inferred models yields a higher number of predicted disease associations. For instance, we found 89 associations unrecorded by DO when requiring them to be present in at least 80% of the models. Exploring the effects of lowering this threshold remains a subject of future research, as we were able to demonstrate our goal to find potentially useful associations using the most stringent threshold. Specifically, two of the fourteen predicted disease-disease associations – between gastric lymphoma and crescentic glomerulonephritis, and between Cushing's syndrome and Hodgkin's lymphoma – demonstrate the ability of the approach to find interesting novel links, but also highlight the fact that it is not possible to determine causal from correlative relationships (which, indeed, in many cases may not be known), given our current scientific understanding.

Perhaps even more interesting is the fact that the newly identified relations between diseases could, in principle, be used to systematically update and extend DO, or even develop a parallel data-driven hierarchy of disease relations. Utilising data fusion for disease re-classification, as well as linking these results with genome-wide association studies (GWAS) is a subject open to future research.

We show that all available molecular data – regardless of their sparseness – are important for effective integration. Surprisingly, we find that genetic interaction data are the most predictive underlying factor of disease-disease associations despite their current small size. The flexibility of our data fusion approach allows us to extend the model with new data sources or omit some sources of information to study their effects on predictive performance. We only require that the underlying graph of data fusion scheme (Figure 3-A) be connected. This gives our data fusion algorithm the power to share latent representations of object types between different data sources. For instance, we cannot omit data on drug targets ($\mathbf{R}_{14}$ in Figure 3-A) without also removing data on adverse side-effects of drug combinations ($\mathbf{\Theta}_4$). Thus, we report in Results on the quality of all models that exclude any reasonable first-order combination of data sources and use these data to estimate contributions of data sources to the quality of the fused model.

Since our data fusion approach is a semi-supervised learning method, it is less prone to over-fitting than supervised methods, i.e. ones that make distinctions between objects on the basis of predefined class label information. Additionally, in order to avoid over-fitting, we selected data fusion parameters through internal cross-validation and used constraint matrices – which express the notion that a pair of similar objects of the same type, such as a pair of drugs or a pair of diseases, should be close in their latent component space – to impose penalties on matrix factors. Thus, the observed reduction in model quality when any one of the included data sets is omitted is caused by the exclusion of complementary information provided by the data set rather than by the lack of robustness of the model.

We have seen the role of data fusion in successful retrieval of existing and uncovering of novel links between diseases. Future improvements of such a comprehensive integration of molecular data would allow better understanding of underlying mechanisms

Table 1 | Data sources. All data sources used in this disease association study, their size, and edge density. Relation matrices R_{ij} relate objects of two different types and their numbers are reported separately (delimited by a forward slash)

Matrix	Data description	# Nodes	# Edges	Density	Reference
$\Theta_1^{(1)}$	Protein-protein interactions	10,360	55,787	0.00104	BioGRID v3.1.94[51]
$\Theta_1^{(2)}$	Gene co-expression	539	869	0.006	Prieto et al.[52]
$\Theta_1^{(3)}$	Cell signalling data	1,217	7,517	0.01016	KEGG[53]
$\Theta_1^{(4)}$	Genetic interactions	542	511	0.00349	BioGRID v3.1.94[51]
$\Theta_1^{(5)}$	Metabolic network	5,908	1,505,831	0.0863	KEGG[53]
Θ_4	Drug interaction data	4,477	21,821	0.00218	DrugBank v3.0[54]
Θ_3	GO semantic structure	11,853	43,924	0.00063	Gene Ontology[28]
Θ_2	DO semantic structure	1,536	1,098	0.00093	Disease Ontology[1]
R_{13}	Gene annotations	17,428/11,853	100,685	0.00049	Gene Ontology[28]
R_{14}	Drug-target relationships	1,978/4,477	7,977	0.00009	DrugBank v3.0[54]
R_{12}	Gene-disease relationships	5,267/1,536	22,084	0.00273	Mapped GeneRIF[55]

that drive diseases and would, in turn, improve choice of medical therapy.

Methods

Data sources. In this study, we integrate biological data on objects of four different types (nodes in Figure 3-A): genes, diseases (Disease Ontology terms), drugs, and Gene Ontology (GO) terms. We observe them through 11 sources of information (edges in Figure 3-A). Every source of information is represented by a distinct data matrix that either relates objects of two different types (such as drugs and their associated target proteins) or objects of the same type (such as genetic interactions between genes): relations between objects of types i and j are represented by a *relation matrix*, $\mathbf{R}_{ij}$, and relations between objects of the same type i are represented by a *constraint matrix*, $\mathbf{\Theta}_i$. Table 1 summarises all 11 data sets.

Disease data. The principal source of information on human disease associations is Disease Ontology (DO)[1]. DO semantically combines medical and disease vocabularies and addresses the complexity of disease nomenclature through extensive cross-mapping of DO terms to standard clinical and medical terminologies of MeSH, ICD, NCI's thesaurus, SNOMED and OMIM. It is designed to reflect the current knowledge of human diseases and their associations with phenotype, environment and genetics. We extract 1,536 DO terms from the latest version of the disease ontology hosted by the OBO Foundry (http://www.obofoundry.org) and construct a binary matrix $\mathbf{R}_{12}$ from 22,084 associations between genes and diseases. DO leverages the semantic richness through linking terms by computable relationships in the hierarchy (e.g. mediastinum ganglioneuroblastoma *is_a* peripheral nervous system ganglioneuroblastoma, which *is_a* ganglioneuroblastoma and then in turn *is_a* neuroblastoma) first by etiology and then by the affected body system. We use the semantic structure of DO to reason over *is_a* relations. Since entries in the constraint matrices are positive for objects that are not similar and negative for objects that are similar, the constraint between two DO terms in $\mathbf{\Theta}_2$ is set to -0.8^{hops}, where hops is the length of the path between corresponding terms in DO graph. We empirically chose 0.8 from [0, 1] range – 0 meaning that no two terms in the DO graph are related, and 1 meaning that two DO terms are always related (regardless of the path distance between them in the DO graph) – by performing standardised internal cross-validation using values between 0 and 1 with a 0.1 step (i.e. 0, 0.1, 0.2, …, 1). Scores of multiple parentage (multiple *is_a* relationships) are summed to produce the final value of semantic association. Throughout the paper, we use *disease* and *DO term* interchangeably, which both refer to a unique DO identifier (DOID).

Gene ontology data. We use relations between 11,853 distinct genes and 100,685 gene annotations that are given by Gene Ontology (GO)[28] to construct a binary matrix of direct annotations $\mathbf{R}_{13}$. Topology of the GO graph is included by reasoning over *is_a*, *part_of* and *has_part* relations between GO terms to populate $\mathbf{\Theta}_3$ in the same way as $\mathbf{\Theta}_2$ with the constraint between two GO terms set to -0.9^{hops}.

Drug data. We obtain drug data from DrugCard entries in the DrugBank (http://www.drugbank.ca) database that contains chemical, pharmacological and pharmaceutical drug information with comprehensive drug target details. Our model contains 4,477 distinct drugs, each identified by a DrugBank accession number. Drugs are related to their target proteins in $\mathbf{R}_{14}$, which is populated by 7,977 binary drug-target relationships from DrugBank. We use reported side-effects of drug combinations form DrugBank as 21,821 binary indicators of interactions between drugs in $\mathbf{\Theta}_4$.

Gene interaction data. We obtain the relationships between genes from five sources of interaction data (top five rows in Table 1). Genes are identified by their NCBI gene IDs. We first map the approved gene symbols and Uniprot IDs to Entrez gene IDs using the index files from HGNC database[29], downloaded in November 2012. This is done to convert all gene annotations, drug-target, and co-expression data into NCBI IDs. To increase coverage of gene and protein interaction data, we include all genes (or equivalently, proteins) for which at least two supporting pieces of information were available in any of the data sources listed in Table 1. In total, these sources include: 55,787 protein-protein interactions (PPIs) between 10,360 proteins ($\mathbf{\Theta}_1^{(1)}$), 869 pairs of co-expressed genes ($\mathbf{\Theta}_1^{(2)}$), 7,517 cell signalling interactions ($\mathbf{\Theta}_1^{(3)}$), 511 human and interspecies genetic interactions ($\mathbf{\Theta}_1^{(4)}$), and 1,505,831 pairs of genes involved in metabolic pathways ($\mathbf{\Theta}_1^{(5)}$).

Data fusion by matrix factorisation. We infer human disease-disease associations by integrating a multitude of relevant molecular data sources. We use a data mining approach based on matrix representation of these molecular data, which works by simultaneous matrix tri-factorisation[24] with sharing of matrix factors. The fusion consists of three main steps (illustrated in Figure 1-A). First, we construct relation and constraint matrices from all available data (Figure 3-A). Recall that a relation matrix encodes relations between objects of two different types (e.g. gene to Gene Ontology term annotation) and a constraint matrix describes relations between objects of the same type (e.g. protein-protein interactions). Then, we simultaneously factorise the relation matrices under given constraints, and finally we score statistically significant associations in the matrix decomposition and identify disease classes (details below and in Žitnik & Zupan (2013)[24]).

Approximate matrix factorisation estimates data matrix $\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$ as a product of low rank matrix factors, $\mathbf{R}_{ij} \approx \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T$, found by solving an optimisation problem. Here, matrix factors are $\mathbf{G}_i \in \mathbb{R}^{n_i \times k_i}$, $\mathbf{S}_{ij} \in \mathbb{R}^{k_i \times k_j}$ and $\mathbf{G}_j \in \mathbb{R}^{n_j \times k_j}$. Factorisation ranks k_i and k_j are chosen to be smaller than both n_i and n_j ($k_i \ll n_i$ and $k_j \ll n_j$), which results in the compressed version of the original matrix $\mathbf{R}_{ij}$. Profiles (row vectors in $\mathbf{R}_{ij}$) of many objects of type i are represented by relatively few vectors from $\mathbf{S}_{ij}$ and low dimensional vectors in $\mathbf{G}_i$ and $\mathbf{G}_j$. Therefore, a good approximation can only be estimated if these vectors span a space that reveals some latent structure present in the original data. The key idea of our data fusion approach is matrix factor sharing when we simultaneously decompose all relation matrices. Matrix factor $\mathbf{G}_i$ is shared across decompositions of relation matrices that relate objects of type i to objects of some other type, whereas $\mathbf{S}_{ij}$ is used only in decomposing $\mathbf{R}_{ij}$. Factor $\mathbf{S}_{ij}$ in our factorised system is thus specific for a relation matrix $\mathbf{R}_{ij}$ and factor $\mathbf{G}_i$ is specific for object type i. They capture source- and object type-specific patterns, respectively.

The objective function minimised by the fusion algorithm enforces a good approximation of the input matrices and is regularised by using available constraint matrices presented in $\mathbf{\Theta}^{(t)}$:

$$\min_{\mathbf{G}\geq 0} \left\|\mathbf{R}-\mathbf{G}\mathbf{S}\mathbf{G}^T\right\|^2 + \sum_{t=1}^{5} \mathrm{tr}\left(\mathbf{G}^T\mathbf{\Theta}^{(t)}\mathbf{G}\right), \tag{1}$$

where $\|\cdot\|$ and $\mathrm{tr}(\cdot)$ denote Frobenius norm and trace, respectively (they are commonly used in matrix approximation tasks). Input to our data fusion algorithm consists of five constraint block matrices $\mathbf{\Theta}^{(t)}$, $1 \leq t \leq 5$ due to five sources of interaction data that represent relations between genes, and a relation block matrix $\mathbf{R}$:

$$\mathbf{\Theta}^{(t)} = \begin{bmatrix} \mathbf{\Theta}_1^{(t)} & 0 & 0 & 0 \\ 0 & \mathbf{\Theta}_2 & 0 & 0 \\ 0 & 0 & \mathbf{\Theta}_3 & 0 \\ 0 & 0 & 0 & \mathbf{\Theta}_4 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 0 & \mathbf{R}_{12} & \mathbf{R}_{13} & \mathbf{R}_{14} \\ \mathbf{R}_{21} & 0 & 0 & 0 \\ \mathbf{R}_{31} & 0 & 0 & 0 \\ \mathbf{R}_{41} & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

The second, third and fourth block along the main diagonal of $\mathbf{\Theta}^{(t)}$ is zero for $t > 1$ because we have a single constraint matrix per disease, drug, and GO term object types. To avoid data redundancy we encode only explicit relations between objects. Such representation leads to zero off-diagonal blocks in $\mathbf{R}$ instead of relation matrices $\mathbf{R}_{23}$, $\mathbf{R}_{24}$, $\mathbf{R}_{32}$, $\mathbf{R}_{34}$, $\mathbf{R}_{42}$ and $\mathbf{R}_{43}$ and to symmetry of relation matrices ($\mathbf{R}_{ji} = \mathbf{R}_{ij}^T$, $\mathbf{S}_{ji} = \mathbf{S}_{ij}^T$). The notion of transitivity between relations is inherently considered by fusion algorithm.

Table 2 | 14 predicted disease-disease associations currently not captured by the semantic structure of Disease Ontology. Literature support for them is listed under the column denoted by ''References''. Reported *p*-values measure how likely it would be for a disease association to emerge if gene-disease relation matrix was permuted, as described in Methods

Disease pair	Literature evidence (quoted verbatim from the referenced source)	References	P-value
vitamin B deficiency (DOID:8449) endogenous depression (DOID:1595)	''Vitamin B complex deficiency causes the psychiatric symptoms of atypical endogenous depression. Dementia and depression have been association with this deficiency possibly from under production of methionine.''	32,33	<0.001
gastric lymphoma (DOID:10540) crescentic glomerulonephritis (DOID:13139)	''Mixed cryoglobulinemia-associated membranoproliferative glomerulonephritis disclosed gastric MALT lymphoma. Glomerulonephritis and lymphoma tend to co-exist in the same patients (relative risk 34.0; $P <$ 0.0001).''	34–36	<0.001
thyroid medullary carcinoma (DOID:3973) cholestasis (DOID:13580)	''Paraneoplastic cholestasis and hypercoagulability associated with medullary thyroid carcinoma. Cholestasis is likely a paraneoplastic effect of thyroid medullary carcinoma.''	37	0.001
crescentic glomerulonephritis (DOID:13139) miliary tuberculosis (DOID:9861)	''Complex-mediated diffuse proliferative glomerulonephritis with crescentic formation is associated with miliary tuberculosis. Antituberculous agents successfully treat miliary tuberculosis and recovered renal function.''	38,39	0.001
thyroid adenoma (DOID:2891) thymoma (DOID:3275)	''Coexistence of bilateral paraganglioma of the A. carotis, thymoma and thyroid adenoma. A common neuroectodermal origin is proposed as an explanation for the coexistence of the carotid body tumor and multiple endocrine tumors.''	40	0.001
early myoclonic encephalopathy (DOID:308) Angelman syndrome (DOID:1932)	''Angelman syndromes share a range of clinical characteristics, including intellectual disability with or without regression and infantile encephalopathy. It presents in infancy with nonspecific features, such as psychomotor delay and seizures. This can lead to the descriptive labels of cerebral palsy or static encephalopathy.''	41,42	<0.001
autoimmune polyendocrine syndrome (DOID:14040) myositis (DOID:633)	''Autoimmune polyendocrine syndrome type 2 (known as Schmidt's syndrome) can be associated with interstitial myositis, an inflammatory myopathy which can be pathologically distinguished from idiopathic polymyositis and inclusion body myositis.''	43	<0.001
primary hyperparathyroidism (DOID:11202) sarcoidosis (DOID:11335)	''Primary hyperparathyroidism simulates sarcoidosis. Coexisting primary hyperparathyroidism and sarcoidosis cause increased Angiotensin-converting enzyme and decreased parathyroid hormone and phosphate levels.''	44	<0.001
cerebrotendinous xanthomatosis (DOID:4810) viral hepatitis (DOID:1884)	''Mutations in the sterol 27-hydroxylase gene (CYP27A) cause hepatitis of infancy as well as cerebrotendinous xanthomatosis. Accumulation of cholesterol and cholestanol can lead to the xanthomata, neurodegeneration, cataracts and atherosclerosis that are typical of cerebrotendinous xanthomatosis.''	45	<0.001
lepromatous leprosy (DOID:10887) mental depression (DOID:1596)	''The precipitating causes of relapse in leprosy include mental depression which downgrades immunity. The prevalence of dementia and depression in older leprosy patients is high.''	46	0.001
male infertility (DOID:12336) DiGeorge syndrome (DOID:11198)	''Complex chromosome rearrangements (CCR) are rare structural chromosome aberrations that can be found in patients with phenotypic abnormalities or in phenotypically normal patients presenting infertility. The malsegregation of CCR can lead to partial 10p12.3 to 10p14 deletion, associated with the DiGeorge like phenotype.''	47,48	0.001
Cushing's syndrome (DOID:12252) Hodgkin's lymphoma (DOID:8543)	''Hodgkin's lymphoma is highly responsive to steroids and Cushing's syndrome results from over exposure to corticosteroids, so it could be considered a treatment side effect. However, the co-existence in one patient of Cushing's disease (caused by a tumour in the pituitary) that suppressed the Hodgkin's lymphoma has been reported.''	49	<0.001
crescentic glomerulonephritis (DOID:13139) prostate cancer (DOID:10283)	''There can be two potential causes for the association: 1) that the drugs and treatment regimen that cancer patients are on causes the glomerulonephritis, or 2) that features of the cancer may cause the glomerulonephritis with ANCA being associated in both cases.''	36	<0.001
allergic bronchopulmonary aspergillosis (DOID:13166) myopathy (DOID:423)	''Allergic Bronchopulmonary aspergillosis is caused by a fungal disease. Fungal diseases are often treated with triazoles. Drug-induced myopathies are well recognised with triazole class of drugs. The association between these two may therefore be based on the treatment and risk it carries, rather than a common mechanism.''	50	<0.001

Data fusion algorithm outputs the block matrix factors **G** and **S**, which we use to identify disease classes:

$$\mathbf{G}=\begin{bmatrix}\mathbf{G}_1 & 0 & 0 & 0\\ 0 & \mathbf{G}_2 & 0 & 0\\ 0 & 0 & \mathbf{G}_3 & 0\\ 0 & 0 & 0 & \mathbf{G}_4\end{bmatrix},\ \mathbf{S}=\begin{bmatrix}0 & \mathbf{S}_{12} & \mathbf{S}_{13} & \mathbf{S}_{14}\\ \mathbf{S}_{21} & 0 & 0 & 0\\ \mathbf{S}_{31} & 0 & 0 & 0\\ \mathbf{S}_{41} & 0 & 0 & 0\end{bmatrix}$$

Notice that each block of matrix **R** is simultaneously approximated as $\mathbf{R}_{ij} \approx \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T$, such that factor $\mathbf{G}_i$ ($\mathbf{G}_j$) is shared among all matrices that relate objects of *i*-th (*j*-th) type to any other object type. That is different from treating **R** as a single homogeneous data matrix, which performs poorly[24].

Parameters of the fusion algorithm are factorisation ranks, k_i, which determine the degree of dimension reduction for four object types in our fusion schema. These factorisation ranks are selected from a predefined set of possible values to optimise the quality of the model in its ability to reconstruct the input data from gene-disease relation matrix $\mathbf{R}_{12}$. For example, gene-disease profiles of length ≈1, 500 in the original space are reduced to profiles with ≈70 factors in data fusion space. We find this approach to be robust and small variations in initial parameter tuning do not impede the overall final quality of the fused system (data not shown). In our study, factorisation ranks of 50 to 80 yield models of similar quality. In general, we find that

Table 3 | Relative contribution of each data source to the fused model. Starting from the configuration given in Figure 3-A, we remove individual data sources, re-run the data fusion algorithm, and compute residual sum of squares (RSS) and explained variance (Evar) changes for the resulting model. For example, if we remove protein-protein interaction data (column labelled "$\mathbf{\Theta}_1^{(1)}$"), the quality of the resulting fused model drops by 2.0% (i.e. RSS increases by 2.0% and Evar decreases by 2.0%). The column labelled "$\Theta_4 + R_{14}$" corresponds to the configuration in which we remove all drug-related information from the system, while the one labelled "Θ_4" indicates that only drug side-effects information was removed

Data source	$\mathbf{\Theta}_1^{(4)}$	$\mathbf{\Theta}_1^{(2)}$	$\mathbf{\Theta}_1^{(3)}$	$\mathbf{\Theta}_1^{(5)}$	$\mathbf{\Theta}_1^{(1)}$	Θ_4	$\Theta_4 + R_{14}$	Θ_3	$\Theta_3 + R_{13}$
RSS increase (↑)	13.3%	6.3%	2.0%	2.0%	2.0%	2.2%	3.8%	1.0%	1.9%
Evar decrease (↓)	9.5%	4.5%	2.5%	2.0%	2.0%	1.3%	4.6%	1.8%	3.2%

if the data contain meaningful information (as opposed to randomised input), the optimised factorisation ranks are much smaller than input dimensions because these data can be effectively compressed, and low-dimensional representation will provide a good estimate of the target relation matrix. Conversely, this would not hold true if we were to predict arbitrarily assigned labels. In that case factorisation ranks would have to be substantially larger in order to produce somewhat comparable models. See Žitnik & Zupan (2013)[24] for a detailed explanation of the algorithm.

Disease class assignment. Each factorisation run produces a set of matrix factors that reconstruct the three relation matrices in our model. For disease association discovery, we are interested in approximating $\mathbf{R}_{12} \approx \mathbf{G}_1\mathbf{S}_{12}\mathbf{G}_2^T$, specifically factor $\mathbf{G}_2$ that contains meta profiles of DO terms and is used to identify classes of diseases. Class membership of a disease is determined by maximum column-coefficient in the corresponding row of $\mathbf{G}_2$. This is a well-known approach in applications of non-negative matrix factorisation[30,31]. A binary connectivity matrix $\mathbf{C}$ is then obtained from class assignments with $\mathbf{C}_{ij}$ set to 1 if disease i and disease j belong to the same class (see algorithm in Figure 1-B). Repeating factorisation process 15 times with different initial random conditions and factorisation ranks gives a collection of connectivity matrices, $\mathbf{C}^{(i)}$, $i \in 1, 2, \ldots, 15$. These are averaged to obtain the consensus matrix $\bar{\mathbf{C}}$ that is then used to assess reliability and robustness of disease associations. The entries in the consensus matrix range from 0 to 1 and indicate the probability that diseases i and j cluster together. If the assignment of diseases into classes is stable, we would expect that the connectivity matrix does not vary among runs and that the entries in the consensus matrix tend to be close to 0 (no association) or to 1 (full consensus for association). To recover informative and relevant disease associations, we are interested in diseases with high values in the consensus matrix. The process is outlined in the algorithm given in Figure 1-B.

Disease associations scoring. Disease associations are scored by permuting the entries in gene-disease relation matrix $\mathbf{R}_{12}$ and inferring the prediction model from the permuted matrix. Matrix $\mathbf{R}_{12}$ encodes relations between genes and diseases, and via genes to the rest of the fusion model, so permuting its entries is sufficient for a complete rewiring of associations. To compute the p-values for the disease associations observed in our inferred model, we generate 70 consensus matrices (each one is averaged over 15 permutations of a disease-gene connectivity matrix, giving $70 \times 15 = 1{,}050$ unique matrices) and express the p-value of a particular disease association as the fraction of factorisation runs in which it was observed.

1. Schriml, L. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* **40**, D940–D946 (2012).
2. Nelson, S., Schopen, M., Savage, A., Schulman, J. & Arluk, N. The MeSH translation maintenance system: structure, interface design, and implementation. *Medinfo* **11**, 67–69 (2004).
3. Aymé, S., Rath, A. & Bellet, B. WHO international classification of diseases (ICD) revision process: incorporating rare diseases into the classification scheme: state of art. *Orphanet J. Rare Dis.* **5**, P1 (2010).
4. Cornet, R. & De Keizer, N. Forty years of SNOMED: a literature review. *BMC Med. Inform. Decis. Mak.* **8**, S2 (2008).
5. Sioutos, N. *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40**, 30–43 (2007).
6. Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for online mendelian inheritance in man (OMIM). *Hum. Mutat.* **32**, 564–567 (2011).
7. Loscalzo, J., Kohane, I. & Barabási, A.-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.* **3**, 124 (2007).
8. Gulbahce, N. *et al.* Viral perturbations of host networks reflect disease etiology. *PLoS Comput. Biol.* **8**, e1002531 (2012).
9. Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA* **105**, 9880–5 (2008).
10. Goh, K.-i. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
11. Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y. & Delisi, C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.* **10**, R91 (2009).
12. Janjić, V. & Pržulj, N. Biological function through network topology: a survey of the human diseasome. *Brief Funct. Genomics* (2012).
13. Emmert-Streib, F., Tripathi, S., Simoes, R., Hawwa, A. & Dehmer, M. The human disease network: opportunities for classification, diagnosis and prediction of disorders and disease genes. *Syst. Biomed.* **1**, 15–22 (2013).
14. Piro, R. M. & Di Cunto, F. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.* **279**, 678–96 (2012).
15. Yu, S. *et al.* L_2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics* **11**, 309 (2010).
16. Mostafavi, S. & Morris, Q. Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics* **12**, 1687–96 (2012).
17. Pandey, G. *et al.* An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput. Biol.* **6** (2010).
18. Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).
19. Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y. & De Moor, B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* **22**, e184–90 (2006).
20. van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., Reinders, M. J. T. & Wessels, L. F. A. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One* **7**, e40358 (2012).
21. De Bie, T., Tranchevent, L.-C., van Oeffelen, L. M. M. & Moreau, Y. Kernel-based data fusion for gene prioritization. *Bioinformatics* **23**, i125–32 (2007).
22. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat. Biotechnol.* **24**, 537–544 (2006).
23. Chen, Z. & Zhang, W. Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLoS Comput. Biol.* **9**, e1002956 (2013).
24. Žitnik, M. & Zupan, B. Data fusion by matrix factorization. (submitted) *Preprint available at Arxiv:1307.0803.* (2013).
25. Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).
26. Holst, F. G. E. *et al.* Low levels of fibrin-stabilizing factor (factor XIII) in human Plasmodium falciparum malaria: correlation with clinical severity. *Am. J. Trop. Med. Hyg.* **60**, 99–104 (1999).
27. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145**, 30–8 (2011).
28. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
29. Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W. & Bruford, E. A. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.* **39**, 514–519 (2011).
30. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
31. Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* **13**, 1706–1718 (2003).
32. Keuter, E. J. [Vitamin B complex deficiency causing the psychiatric symptoms of atypical endogenous depression]. *Ned. Tijdschr. Geneeskd.* **102**, 1501–1503 (1958).
33. Carney, M. *et al.* Red cell folate concentrations in psychiatric patients. *J. Affect. Disorders* **19**, 207–213 (1990).
34. Buob, D. & Copin, M. C. [Mixed cryoglobulinemia-associated membranoproliferative glomerulonephritis, disclosing gastric MALT lymphoma]. *Ann. Pathol.* **26**, 267–270 (2006).
35. Skopouli, F. N., Dafni, U., Ioannidis, J. P. & Moutsopoulos, H. M. Clinical evolution, and morbidity and mortality of primary Sjögren's syndrome. *Semin. Arthritis Rheu.* **29**, 296–304 (2000).
36. Von Vietinghoff, S., Schneider, W., Luft, F. & Kettritz, R. Crescentic glomerulonephritis and malignancy-guilty or guilt by association? *Nephrol. Dial. Transpl.* **21**, 3324–3326 (2006).
37. Tiede, D. J., Tefferi, A., Kochhar, R., Thompson, G. B. & Hay, I. D. Paraneoplastic cholestasis and hypercoagulability associated with medullary thyroid carcinoma. Resolution with tumor debulking. *Cancer* **73**, 702–705 (1994).

38. Kohler, L. J., Gohara, A. F., Hamilton, R. W. & Reeves, R. S. Crescentic fibrillary glomerulonephritis associated with intermittent rifampin therapy for pulmonary tuberculosis. *Clin. Nephrol.* **42**, 263–265 (1994).
39. Wen, Y. K. & Chen, M. L. Crescentic glomerulonephritis associated with miliary tuberculosis. *Clin. Nephrol.* **71**, 310–313 (2009).
40. Refior, M. & Mees, K. Coexistence of bilateral paraganglioma of the A. carotis, thymoma and thyroid adenoma: a chance finding? *Laryngorhinootologie* **79**, 337–340 (2000).
41. Willemsen, M. H., Rensen, J. H., van Schrojenstein-Lantman de Valk, H. M., Hamel, B. C. & Kleefstra, T. Adult phenotypes in Angelman- and Rett-like syndromes. *Mol. Syndromol.* **2**, 217–234 (2012).
42. Dagli, A., Buiting, K. & Williams, C. A. Molecular and clinical aspects of Angelman syndrome. *Mol. Syndromol.* **2**, 100–112 (2012).
43. Heuss, D., Engelhardt, A., Gobel, H. & Neundorfer, B. Myopathological findings in interstitial myositis in type II polyendocrine autoimmune syndrome (Schmidt's syndrome). *Neurol. Res.* **17**, 233–237 (1995).
44. Lim, V. & Clarke, B. L. Coexisting primary hyperparathyroidism and sarcoidosis cause increased Angiotensin-converting enzyme and decreased parathyroid hormone and phosphate levels. *J. Clin. Endocr. Metab.* **98**, 1939–1945 (2013).
45. Clayton, P. T. *et al.* Mutations in the sterol 27-hydroxylase gene (CYP27A) cause hepatitis of infancy as well as cerebrotendinous xanthomatosis. *J. Inherit. Metab. Dis.* **25**, 501–513 (2002).
46. Su, T. W., Wu, L. L. & Lin, C. P. The prevalence of dementia and depression in Taiwanese institutionalized leprosy patients, and the effectiveness evaluation of reminiscence therapy–a longitudinal, single-blind, randomized control study. *Int. J. Geriatr. Psychiatry* **27**, 187–196 (2012).
47. Karmous-Benailly, H. *et al.* Unbalanced inherited complex chromosome rearrangement involving chromosome 8, 10, 11 and 16 in a patient with congenital malformations and delayed development. *Eur. J. Med. Genet.* **49**, 431–438 (2006).
48. Christopoulou, G. *et al.* Clinical and molecular description of the prenatal diagnosis of a fetus with a maternally inherited microduplication 22q11.2 of 2.5 Mb. *Gene* **527**, 694–697 (2013).
49. Howell, D., Bergsagel, J., Chu, R. & Meacham, L. Suppression of Hodgkin's disease in a patient with Cushing's syndrome. *J. Pediatr. Hematol. Oncol.* **26**, 301–303 (2004).
50. Valiyil, R. & Christopher-Stine, L. Drug-related myopathies of which the clinician should be aware. *Curr. Rheumatol. Rep.* **12**, 213–220 (2010).
51. Stark, C. *et al.* The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* **39**, 698–704 (2011).
52. Prieto, C., Risueño, A., Fontanillo, C. & De Las Rivas, J. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One* **3**, e3911 (2008).
53. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
54. Knox, C. *et al.* Drugbank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* **39**, 1035–1041 (2011).
55. Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10**, S6 (2009).

Acknowledgments

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, NIH programme grant P01 HD39691, GlaxoSmithKline (GSK) Research and Development Ltd, the Slovenian Research Agency Programme Grant P2-0209, EU FP7 project "CARE-MI" (Health-F5-2010-242038), ARRS project J1-5454, and the Serbian Ministry of Education and Science Project III44006.

Author contributions

M.Z., V.J., C.L., B.Z. and N.P. designed the experiments. M.Z. performed the experiments. M.Z., V.J., C.L., B.Z. and N.P. wrote the main manuscript text. All authors reviewed the manuscript. The authors have no competing financial interests.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Žitnik, M., Janjić, V., Larminie, C., Zupan, B. & Pržulj, N. Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* **3**, 3202; DOI:10.1038/srep03202 (2013).