Integration of molecular network data reconstructs Gene Ontology
Vladimir Gligorijević, Vuk Janjić and Natasa Pržulj*
Department of Computing, Imperial College London SW7 2AZ, UK

ABSTRACT
Motivation: Recently, a shift was made from using Gene Ontology (GO) to evaluate molecular network data to using these data to construct and evaluate GO. Dutkowski et al. provide the first evidence that a large part of GO can be reconstructed solely from topologies of molecular networks. Motivated by this work, we develop a novel data integration framework that integrates multiple types of molecular network data to reconstruct and update GO. We ask how much of GO can be recovered by integrating various molecular interaction data.

Results: We introduce a computational framework for integration of various biological networks using penalized non-negative matrix tri-factorization (PNMTF). It takes all network data in a matrix form and performs simultaneous clustering of genes and GO terms, inducing new relations between genes and GO terms (annotations) and between GO terms themselves. To improve the accuracy of our predicted relations, we extend the integration methodology to include additional topological information represented as the similarity in wiring around non-interacting genes. Surprisingly, by integrating topologies of bakers’ yeasts protein–protein interaction, genetic interaction (GI) and co-expression networks, our method reports as related 96% of GO terms that are directly related in GO. The inclusion of the wiring similarity of non-interacting genes contributes 6% to this large GO term association capture. Furthermore, we use our method to infer new relationships between GO terms solely from the topologies of these networks and validate 44% of our predictions in the literature. In addition, our integration method reproduces 48% of cellular component, 41% of molecular function and 41% of biological process GO terms, outperforming the previous method in the former two domains of GO. Finally, we predict new GO annotations of yeast genes and validate our predictions through GIs profiling.

Availability and implementation: Supplementary Tables of new GO term associations and predicted gene annotations are available at http://bio-nets.doc.ic.ac.uk/GO-Reconstruction/.
Contact: natasha@imperial.ac.uk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
In many areas of biomedical research, ontologies play an important role in unification of knowledge as a hierarchy of terms and their mutual relationships. Among widely used ontologies is Gene Ontology (GO), which describes genes and gene products in terms of their associated biological process (BP), molecular function (MF) and cellular component (CC) (Ashburner et al., 2000). GO is a current major source of information for annotating genes and proteins across various species and providing tools for systematic assessment of experimental gene sets via enrichment analysis.

Since its foundation, GO has been growing in size and complexity containing today vast amounts of annotated biological data. Initially, GO was manually curated by domain experts and members of the research and annotation communities. However, because of their inconsistency in translation to GO terms and relations, manual curations have encountered many difficulties (Ashburner et al., 2001). Additionally, rapid development of technologies for biological data acquisition has resulted in an accumulation of biological data exceeding our ability to interpret (Chen and Xu, 2004).

To overcome these problems, many computational tools for automatic gene and protein annotation have been devised. Much effort has been invested in assessing the accuracy of such annotation predictions (Radojovic et al., 2013). Methods for gene annotation prediction have either followed approaches that transfer annotations from well-observed to partially observed genes based solely on sequence similarity (Loewenstein et al., 2009) or approaches that directly predict function of unknown genes using machine learning methods (Clare and King, 2003). Recent methodologies focus more on integration of distinct biological data sources, which contribute to more accurate predictions of gene annotation.

The availability of genomic-level information from high-throughput measurements of genetic and protein interactions, messenger RNA expression profiles and metabolic pathways has created new opportunities for function prediction. A major challenge is how to integrate all these diverse data to predict annotations of yet unannotated proteins. Among the widely used computational methods addressing this problem are Bayesian reasoning (Chen and Xu, 2004), network-based analysis (Mostafavi and Morris, 2010; Mostafavi et al., 2008), kernel-based statistical learning (Lanckriet et al., 2004) and matrix factorization-based data fusion (Žitnik and Blaz, 2014). All these methods have demonstrated that the integration of complementary biological data significantly improves accuracy of gene function annotation prediction.

Recent work incorporated large gene and protein interaction networks into a probabilistic clustering procedure to reconstruct the GO (Dutkowski et al., 2013). It identified new terms and relations that were missing from GO based solely on network topology. This work provides evidence that a large part of GO can be reconstructed using only topologies of molecular networks.

In this work, we propose a new data integration method for prediction of GO term annotations of unannotated genes and finding new relations between existing GO terms purely from network topology. The method is based on penalized non-negative matrix tri-factorization (PNMTF) for heterogeneous
data clustering (Wang et al., 2008, 2011), PNMTF has been used for prediction of disease associations (Zitnik et al., 2013), identification of cancer subtypes (Liu et al., 2014), predicting protein–protein interactions (PPIs) (Wang et al., 2013) and detecting phenotype–gene associations (Hwang et al., 2012).

Here, we extend this method to take multiple types of molecular network data and use them to reconstruct and update GO with new information. We apply our method to Saccharomyces cerevisiae data used by Dutkowski et al. (2013): PPI network, genetic interaction (GI) network, gene co-expression (Co-Ex) network and integrated functional network known as YeastNet (Lee et al., 2007).

Our method takes all data in a matrix form and performs simultaneous clustering of genes and GO terms inducing new associations between genes and GO terms and between GO terms themselves. We extend the integration methodology to include similarity in wiring around non-interacting genes. We measure this by distance graphlet degree vectors (GDVs) (Pržulj, 2007). Graphlets and graphlet-based measures have bridged molecular network topology and biological function. For instance, simple homogeneous clustering of proteins in a PPI network based on the GDV similarity has revealed groups of proteins with a common biological function (Milenković and Pržulj, 2008; Milenković et al., 2010).

Therefore, we add these to incorporate more topology into the integration process and improve accuracy of predictions. Using various measures for assessing the quality of our prediction, we systematically examined the contribution of these additional topological constraints to GO prediction. Graphlet-based similarity has not been exploited in any of the previous network integration approaches.

Surprisingly, we find that our method can successfully reconstruct almost the entire GO by using solely topology of molecular interaction networks. Furthermore, we predict new GO term associations and gene annotations from integrated topologies of molecular interaction network and validate our predictions.

2 METHODS

2.1 Matrix tri-factorization for data integration

We use a co-clustering algorithm based on PNMTF to integrate multi-type biological data. The clustering analysis is used to infer new relations between data objects that were not previously present in the data. Such a technique makes use of all available information presented in the network form, including both inter-type relations and intra-type constraints (Ding et al., 2006; Wang et al., 2008). This algorithm aims to simultaneously cluster data using the interrelatedness between data types under the guidance of some prior knowledge given in the form of intra-type pairwise constraints. These constraints often indicate similarity or dissimilarity relationships between data objects of the same type. Constraints guide the clustering procedure so that similar objects can belong to the same cluster while dissimilar cannot.

The simplest co-clustering problem involves only two types of objects (e.g. genes and GO terms) with size $n_1$ and $n_2$. If there are $n_1$ objects of the first type and $n_2$ objects of the second type, then we have an inter-type relationship matrix $R_{12} \in \mathbb{R}^{n_1 \times n_2}$ with an entry $R_{12}(i,j)$ representing the relationship between $i$-th data point in the first dataset and the $j$-th data point in the second dataset. Simultaneous clustering of these datasets can be seen as a solution of the non-negative matrix tri-factorization (NMTF) problem where a given relation matrix, $R_{12} \in \mathbb{R}^{n_1 \times n_2}$ can be approximated as the product of three low-rank matrix factors:

$$R_{12} \approx G_1 S_{12} G_2^T,$$

where non-negative $G_1 \in \mathbb{R}^{n_1 \times k_1}$ and $G_2 \in \mathbb{R}^{n_2 \times k_2}$ correspond to the cluster indicator matrix of the first and the second dataset, and $S_{12} \in \mathbb{R}^{k_1 \times k_2}$ corresponds to compressed low-dimensional version of the initial relation matrix. Rank factors, $k_1$ and $k_2$, are often chosen to be much smaller than the corresponding matrix dimensions ($k_1 \ll n_1$, $k_2 \ll n_2$). NMTF algorithm minimizes the following objective function:

$$\min_{G_1 \geq 0, G_2 \geq 0} J = \| R_{12} - G_1 S_{12} G_2^T \|^2$$  (1)

This objective function can be further used to incorporate intra-type constraints whose violation causes penalties. Constraints that relate data points, $i$ and $j$, in two different datasets are represented via two constraint matrices, $\Theta_1 \in \mathbb{R}^{n_1 \times n_1}$ and $\Theta_2 \in \mathbb{R}^{n_2 \times n_2}$. Entries of the constraint matrix are positive for dissimilar data objects because they impose penalties on the current approximation given in the Equation (1). Entries of the constraint matrix are negative for similar objects because they are rewords that reduce the objective function. Therefore, the constraint matrices can be included as additional penalty terms in the objective function in the following way:

$$\min_{G_1 \geq 0, G_2 \geq 0} J = \| R_{12} - G_1 S_{12} G_2^T \|^2 + \tau (G_1^T \Theta_1 G_1) + \tau (G_2^T \Theta_2 G_2)$$  (2)

where $\tau$ denotes the trace of a matrix. This optimization problem is known as PNMTF problem. Its solution produces two matrix factors, $G_1$ and $G_2$, that can be interpreted as the cluster indicator matrices for the first and the second dataset. Specifically, factor $G_1$ is used to assign data objects from the first dataset to clusters so that data object $i$ is placed in the cluster $i$ if $G_1(i,i)$ is the largest entry in column $i$ (Brunet et al., 2004). This assignment procedure results in a binary connectivity matrix, $C$, of size $n_1 \times n_1$ with entry $C(p,q) = 1$ if objects $p$ and $q$ belong to the same cluster and $C(p,q) = 0$ otherwise. Hence, an integration of all data sources is achieved by clustering the first and the second datasets simultaneously using $R_{12}$, $\Theta_1$ and $\Theta_2$ that encode the data.

Biological entities, such as genes and proteins engage in various molecular interactions, or are connected through GO relationships. We represent these as networks and integrate their network topology (also called structure) in the form of constraints of the objective function. These constraints are implemented into the objective function in the form of network Laplacians (Hwang et al., 2012; Wang et al., 2011). That is, we are now minimizing:

$$\min_{G_1 \geq 0, G_2 \geq 0} J = \| R_{12} - G_1 S_{12} G_2^T \|^2 + \tau (G_1^T \mathcal{L}_1 G_1) + \tau (G_2^T \mathcal{L}_2 G_2)$$  (3)

where $\mathcal{L}_1 = D_A - A_\alpha$ represents network Laplacian of the molecular network of the $\alpha$ data type; $A_\alpha$ is the network adjacency matrix, and $D_A$ is the diagonal degree matrix with entries being row summation of the matrix $A_\alpha$. $D_A(i,i) = \sum_j A_\alpha(i,j)$. These additional, Laplacian-based terms encourage the connected or ontology-related genes (proteins) in the network to be assigned to the same cluster. To integrate the network data and predict GO term relationships from network topology, along with new gene annotations, the term $\tau (G_1^T \mathcal{L}_1 G_1)$ imposes that interacting genes get placed into the same cluster and similarly $\tau (G_2^T \mathcal{L}_2 G_2)$ imposes that linked GO terms get placed into the same cluster.

2.2 Integration of various constraints on the same objects

Most of the biological datasets include various types of interactions (i.e. constraints) over the same set of entities. For instance, genes might interact via GIs, and they also might be related based on the correlation of
their expression profiles. The former is known as a GI network, and the latter as a gene Co-Ex network. To properly integrate this information into the clustering procedure, we make an improvement to the regularized PNMTF optimization problem. We extend it to take into account multiple constraints over the objects of the same type. Suppose we have a set of $N$ adjacency matrices $\{A_1, A_2, \ldots, A_N\}$, representing $N$ data sources relating objects of the first type. By adding these constraints in the Laplacian form as penalty terms into our objective function (3), we end up with the following:

$$
\min_{G_i \geq 0, G_j \geq 0} J = \| R - G_i S_j G_j^T \|^2 
+ \sum_{\beta=1}^N tr(G_i^T L^\beta G_1) + tr(G_i^T L^\beta G_j)
$$ (4)

Integration of all available information about a particular data type has demonstrated to lead better predictions of new relationships among data objects. For example, the integration of all available human molecular networks yields a successful reproduction of the existing and prediction of new associations between diseases (Žitnik et al., 2013).

Unlike previous works where only network connections are considered as constraints (Hwang et al., 2012; Liu et al., 2014; Wang et al., 2013; Zhang et al., 2011; Žitnik et al., 2013), our approach takes a step further by incorporating additional constraints in the form of topological similarity between nodes in a network that are not necessarily linked. Here, we use the topological similarity measure based on GDVs. Graphlets are small non-isomorphic induced substructures of a large network (Pržulj et al., 2004). There are 29 graphlets containing 2–5 nodes. By taking into account the symmetries between nodes in a graphlet, we can distinguish between 73 automorphic orbits. Counting how many times a particular node touches any of 73 different orbits, we may define a 73-dimensional network, and the Laplacian matrix for a node $u$ is defined as a logarithmic distance between nodes’ orbit memberships:

$$
D(u) = \sum_{i=1}^{73} D_i \frac{w_i}{\sum_{i=1}^{73} w_i},
$$ (5)

where $D_i(u) = \log (w_i + 1) - \log (w_i - 1)$ is assigned to each orbit $i \in \{0, \ldots, 72\}$. The weight, $w_i$, measures to which extent orbit $i$ is affected by other orbits. Higher weights are assigned to orbits that are less affected by other orbits, whereas lower weights are assigned to orbits that are affected by many other orbits. The number of orbits that affect orbit $i$ is given by $g_i$.

Using the distance measure defined in Equation (5), GDV similarity between nodes $u$ and $v$ is measured as

$$
S(u, v) = 1 - D(u, v)
$$

GDV similarity measure has been used for predicting biological function of unclassified proteins (Milenković and Pržulj, 2008), classification of cancer and non-cancer genes (Milenković et al., 2010) and prediction of new cardiovascular disease genes (Sarajlic et al., 2013).

Here, we include GDV similarity measure into our objective function (Equation 4) as followings. For each of the given data source $\beta$ (i.e. biological network), we construct a similarity matrix $S_\beta \in \mathbb{R}^{n \times n}$. Then, by computing a statistically significant threshold for topological similarity of two nodes in each of the GDV similarity matrices, we consider only data objects (genes/proteins) with GDV similarity higher than the computed threshold (see Supplementary Fig. S2):

$$
\Sigma_\beta(u, v) = \begin{cases} 
1, & \text{if } S_\beta(u, v) \geq S^\text{threshold}_\beta \\
0, & \text{if } S_\beta(u, v) < S^\text{threshold}_\beta 
\end{cases}
$$

Topological similarity constraints are again implemented into the objective function through Laplacian regularization:

$$
\min_{G_i \geq 0, G_j \geq 0} J = \| R - G_i S_j G_j^T \|^2 
+ \sum_{\beta=1}^N tr(G_i^T L^\beta G_1) + \sum_{\beta=1}^N tr(G_i^T L^\beta G_j) + tr(G_i^T L^\beta G_j)
$$ (6)

where, $\Lambda = D - \Sigma$ is a Laplacian of $\Sigma$ matrix, and $D$ is the diagonal matrix with entries equal to the row summation of $\Sigma$ matrix.

### 2.3 Multiplicative update algorithm

We extend the original PNMTF optimization (Wang et al., 2008) to handle the additional penalty terms and network regularizations in Equation (6). Solving the optimization problem results in the following multiplicative update rules for matrix factors $G_1$, $G_2$, and $S_1$ (Wang et al., 2008):

$$
S_{12} \leftarrow \langle G_1^T G_1 \rangle^{-1} G_1^T R S_2 G_2^T G_2^{-1}
$$ (7)

$$
G_1(i,j) \leftarrow \frac{(R_{12} G_1 S_{12}^T)^j + [G_1 S_{12}^T G_2 S_{12}^{-1}]^j + [L^\kappa + \Lambda^\kappa] G_1]}{(R_{12} G_1 S_{12}^T)^j + [G_1 S_{12}^T G_2 S_{12}^{-1}]^j + [L^\kappa + \Lambda^\kappa] G_1}
$$ (8)

$$
G_2(i,j) \leftarrow \frac{(R_{12} G_1 S_{12}^T)^j + [G_1 S_{12}^T G_2 S_{12}^{-1}]^j + [L^\kappa + \Lambda^\kappa] G_2]}{(R_{12} G_1 S_{12}^T)^j + [G_1 S_{12}^T G_2 S_{12}^{-1}]^j + [L^\kappa + \Lambda^\kappa] G_2}
$$ (9)

where we use $+$ and $-$ signs in superscripts to denote non-negative matrices $M^+$ and $M^-$ of a matrix $M$, respectively, defined as $M^+ = \frac{M + M^T}{2}$ and $M^- = \frac{M - M^T}{2}$. The algorithm starts by randomly initializing matrices $G_1$ and $G_2$, which are iteratively updated to minimize objective function in Equation (6). The rigorous proof of the correctness and convergence of these update rules can be found in (Wang et al., 2008). Under these update rules, the objective function $J$ [Equation (6)] is guaranteed not to increase. Hence, we look at the change in the objective function between two consecutive iterations and define the stopping criterion as $J_k - J_{k-1} < \epsilon$.

In all our runs, parameter $\epsilon$ is set to $10^{-5}$, which was shown to be significant to minimize the objective function. Compared with the probabilistic clustering approach for GO reconstruction presented by Dukowski et al. (2013), our approach is computationally more demanding because of slow convergence of multiplicative update rules. However, our approach is more general, as it can integrate any number and type of heterogeneous data that could lead to more accurate predictions.

### 2.4 Predicting associations between GO terms

Each factorization run produces matrix factors $G_1$ related to gene set, and $G_2$ related to GO terms. We use $G_2$ factor to construct connectivity matrix $C$ as described in the Section 2.1. Clusters of mutually related GO terms are obtained from the connectivity matrix. To assess reliability and robustness of GO term associations prediction, we use the stochastic property of our algorithm. We perform multiple runs with the same rank parameters and different initial random initializations and construct a set of 20 different connectivity matrices: $\{C^{(1)}, \ldots, C^{(20)}\}$. Then, we compute the consensus matrix $\bar{C}$, defined as the average over all
connectivity matrices. Thereby entries in the consensus matrix range from 0 to 1, and they can be interpreted as probabilities that two GO terms, GO_i and GO_j, belong to the same cluster. To predict new GO term associations, we are only interested in values of probability equal to one because they correspond to the case of hard clustering, in which there is no overlap between clusters, and hence, there is no ambiguity in predicted GO term associations. The complete algorithm for prediction of new GO term association is summarized in Algorithm 1.

Algorithm 1 GO term associations prediction

Input: Relation matrix: R_{ij}; constraint matrices: A^\beta_i, L^\beta_i, for networks \beta \in \{1, 2, 3, 4\} L^\beta_i for Gene Ontology; rank parameters k_1 and k_2

Output: Consensus matrix C

for \(i \in \{1, 20\}\) do
  Initialize \(G_1\) and \(G_2\)
  while not |\(J_{i, L, j} - J_{i, L, j}^{old}\)| < \(\epsilon\) do
    Update \(S_1\) using Equation (7) while keeping fixed \(G_1\) and \(G_2\)
    Update \(G_1\) using Equation (8) while keeping fixed \(G_1\) and \(S_2\)
    Update \(G_2\) using Equation (9) while keeping fixed \(G_1\) and \(S_2\)
  Compute connectivity matrix \(C^{(i)}\) for GO terms using \(G_2\) for class assignment
  Compute the average connectivity matrix as: \(C = \frac{1}{20}C^{(i)}\)

Extract new GO term relations:
\(G = \{GO_i, GO_j\}/\{GO_i, GO_j\} \forall GO_i, GO_j \in \{all \ GO \ terms\} \times \{i, j\} = 1\)

To assess the statistical significance of GO term associations, we compute the P-value in the following way. First, we remove any prior knowledge on GO term relations (i.e. we remove matrix \(L_2\)). Then, we run our algorithm 100 times, each time with different relations matrix obtained by permuting the entries of the original relations matrix, \(R_{12}\). In total, we obtain \(100 \times 20 = 2000\) different connectivity matrices. We define the P-value of a particular GO term association as the fraction of connectivity matrices in which that particular association is observed.

2.5 Rank parameters selection

Input parameters of our algorithm are factorization ranks, \(k_1\) and \(k_2\), which we systematically examine and choose to achieve a correct reduction of dimensionality of our data. These factorization ranks capture the meaningful information that can further be decomposed into clusters.

There is no agreed-upon procedure for choosing the right factorization ranks. The most common approach, widely used in many dimension reduction problems is *cophenetic correlation coefficient*, as a quantitative measure of stability for clustering (Brunet et al., 2004). For a given factorization rank, cophenetic correlation coefficient is computed over the values of the consensus matrix, \(\rho(T)\). It is defined as the Pearson’s correlation coefficient between the distance matrix, \(1 - T\), and the matrix of cophenetic distances obtained by the linkage used in hierarchical clustering for re-ordering \(T\). If the clustering is stable, i.e. the entries in \(T\) are close to 0 or 1, then \(\rho(T) \approx 1\), otherwise, if the entries are scattered between 0 and 1, \(\rho(T) < 1\).

A simple generalization of this procedure applied to two types of our data (genes and GO terms) includes computation of cophenetic correlation coefficient for each of the consensus matrices, \(T_{ep}\) (for genes), \(T_{GO}\) (for GO terms), and then we define the average cophenetic correlation coefficient as

\[
\rho_{avg} = \frac{\rho(T_{ep}) + \rho(T_{GO})}{2}
\]

We search for the values of \(k_1\) and \(k_2\), that maximize \(\rho_{avg}\). We do this by running our algorithm for all \((k_1, k_2)\) pairs such that \(0 < k_1, k_2 < 60\), so that we would capture the best dimensionality of our data (see below).

2.6 Datasets and preprocessing

To make our study directly comparable with the competing method for reconstructing GO from network data, we run our method on the same S.cerevisiae data as Dutkowski et al. (2013); PPI network from BioGRID (Chatr-Aryamontri et al., 2013), GI network from DRYGIN (Costanzo et al., 2010), gene Co-Ex network from SMD (Hubble et al., 2009) and integrated function network, YeastNet, from (Lee et al., 2007). For each of the these networks, we construct Laplacian constraint matrices, \(L_1, L_2, L_3, L_4\), respectively.

To apply multiplicative update rules, we make all data matrices of the same dimension: we construct them over the union of genes presented in all four data sources. The semantic structure of GO is also taken into account in our integration algorithm. We extract all GO terms for S.cerevisiae and create \(L_2\) constraint matrix as follows. First, we construct a directed acyclic ontology graph using the four basic semantic types of GO relations: is_a, part_of, regulates and has_part. Then, we assign value 0.9 to each pair of GO terms as a measure of association strength, where \(l\) is the length of directed shortest path between terms in the ontology graph. This allows us to also take into account mutual influence of hierarchically distant non-adjacent GO terms (Zhu et al., 2005). The value of 0.9 is chosen from empirical observations, as described by Zittnik et al. (2013). Finally, we construct the Laplacian constraint matrix, \(L_2\), by using these values of association strengths.

Annotation files from GO are used to construct the binary relation matrix, \(R_{12}\), with entries \(R_{12}(i, j) = 1\) if gene \(i\) is annotated by GO term \(j\) and 0 otherwise. For each of the aforementioned biological networks, we also compute GDV similarity constraint matrices: \([L_1, L_2, L_3, L_4]\). As we describe in Section 2.2, we only consider gene pairs with statistically significant GDV similarity. All these network data are schematically represented in Figure 1.

3 RESULTS AND DISCUSSION

We apply our algorithm to identify new GO term relations and annotate proteins with existing GO terms by integrating multiple independent network sources given in the Table 1. We find that the optimal rank parameters \(k_1\) and \(k_2\) are 58 and 56, respectively.
(see Supplementary Fig. S3). We examine the contribution of each data source to the integration model.

3.1 Contribution of data to the integration model
We estimate the influence of each network on our integration model by comparing the quality of the initial model (consisting of four networks and their corresponding GDV similarity matrices) with the quality of the model with one data source removed from the initial set. Models are evaluated through residual sum of squares (RSS), \( \text{RSS}(R_{12}) = \sum_{i,j} |R_{12}(i,j) - (G_i S_{12} G_j^T(i,j))^2 | \), and explained variance (Evar), \( \text{Evar}(R_{12}) = 1 - \text{RSS}(R_{12})/\sum_{i,j} |R_{12}(i,j)|^2 \), that measure the performance of the matrix factorization algorithm and its ability to accurately reproduce the gene–GO term relation matrix. Low values of RSS and high values of Evar indicate better quality of the model (Hutchins et al., 2008).

We find that with the removal of each of the four data sources (a network along with its corresponding GDV similarity matrix) the value of RSS increases, while the value of Evar decreases, implying that each data source contributes to the quality of the model. Relative increase of RSS and relative decrease of Evar (with respect to the initial model containing all the data), computed by removing a particular network along with its corresponding GDV similarity matrix, are shown in the top panel of Figure 2. We find that the largest model degradation is achieved with the removal of GI network and its corresponding GDV similarity matrix. A similar result was reported by Žitnik et al. (2013): they found GIs to be the most informative data source in prediction of disease–disease associations. Exclusion of the gene Co-Ex network and its corresponding GDV similarity matrix confirms the observation of Dutkowski et al. (2013) that a combination of YeastNet and Co-Ex network contributes the least to the quality of the model (RSS = 0.8%, Evar = 1%).

3.2 Recovering existing knowledge
Our integration of the biological networks and their corresponding GDV similarities results in a set of highly reliable GO term classes, represented as clusters in a block diagonal form of the consensus matrix. Size distribution of these clusters and the consensus matrix are shown in the Supplementary Figure S4. In addition to this experiment, we also perform the same analysis on the data consisting only of biological networks (excluding GDV similarities from our integration procedure). This allows us to compare the clustering results of different integration models and to estimate the importance of additional topological constraints.

To evaluate the performance of our methodology in reproducing GO term relations, we look at the overlap between cluster members and the existing GO hierarchy and find that on average 92% of cluster members are directly connected via semantic relations in GO. These cluster-induced GO term relations are confirmed to be statistically significant (\( P \leq 0.01 \), computed as explained in Section 2.4). A slightly lower score of 90% is

Table 1. All networks used in this study

| Data               | Matrix representation | Matrix dimension | NNZ* |
|--------------------|-----------------------|------------------|------|
| PPI                | \( \mathcal{L}_4^{(3)} \) | 3401 × 3401      | 26596|
| GI                 | \( \mathcal{L}_4^{(2)} \) | 3090 × 3090      | 22480|
| Co-Ex              | \( \mathcal{L}_4^{(1)} \) | 228 × 228        | 3410 |
| YeastNet           | \( \mathcal{L}_4^{(4)} \) | 3351 × 3351      | 21146|
| GDV similarity (PPI)| \( \Lambda_4^{(1)} \)  | 1609 × 1609      | 93536|
| GDV similarity (GI)| \( \Lambda_4^{(2)} \)  | 1550 × 1550      | 89434|
| GDV similarity (Co-Ex)| \( \Lambda_4^{(3)} \)  | 122 × 122        | 2524 |
| GDV similarity (YeastNet)| \( \Lambda_4^{(4)} \)  | 1453 × 1453      | 88986|
| \( \mathcal{L}_2 \) | GO semantic structure | 3993 × 3993      | 15872|
| \( \mathbf{R}_{12} \) | Gene annotation      | 5051 × 3993      | 45782|

Note: Matrix dimensions are given before unioning genes in all data to obtain the same dimension of matrices (see Section 2.6). GDV matrices are of different dimension than \( \mathcal{L} \)-matrices because they contain only genes that are statistically significantly similar (see Section 2.2).

*Number of non-zero entries in a matrix.

Fig. 2. Relative contribution of each data source to the integration model measured by RSS, blue, and Evar, red. The top panel shows the relative changes in RSS and Evar with the removal of a particular network and its corresponding GDV similarity matrix. The bottom panel shows the same measures but only with removal of GDV similarity matrices.
achieved when considering only the network data without GDV similarity matrices, indicating that graphlet similarity matrices contribute to capturing relations, which would otherwise be missed.

Furthermore, we examine the robustness of this result to the removal of particular datasets. Surprisingly, we find that omission of GDV similarity matrix of gene Co-Ex network contributes the most to the predictive performance of our algorithm, leading to the maximum of 96% of recovered GO terms. Hence, inclusion of GDV similarity of gene Co-Ex network introduces noise into the integration procedure, wrongly guiding the clustering process, which in turn results in lower prediction performance. This is a consequence of the random GDV similarity distribution over all genes in the gene Co-Ex network (Supplementary Fig. S2C). Given that inclusion of GDV similarity matrix of the Co-Ex network impairs the predictive performance of our algorithm and because we have shown that its exclusion makes minimal effect on the quality of the model, we discard that data source from further analysis.

Surprisingly, recovering 96% of GO terms that are directly related in GO (this is not a percentage of recovered relations between GO terms) by is_a, part_of, regulates and has_part associations, indicates that entire GO could, in principle, be reconstructed solely from topologies of molecular interaction networks. Reporting this statistic is consistent with what previous studies using a similar methodology reported (Zitnik et al., 2013). When we say that ‘96% of GO terms is recovered’, we mean that our methodology correctly identifies a set of 96% of GO terms that contain relations between them. This does not mean that this set is fully connected (i.e. that each pair of GO terms in it is related). Our set of 96% of GO terms contains 78% of all relations currently present in GO. To our knowledge, because a large part of GO is sequence derived, this is the first conformation that network topology and sequence carry similar biological information.

To further validate the performance of our methodology in reconstruction of GO terms, we use the gene–GO term relation matrix, reconstructed from matrix factors $R_{12} = G_{1} S_{12} G_{2}^{T}$. Its entries indicate the annotation strength of a gene, $i$, related to a GO term $j$, with $R_{12}(i,j) = 0$ denoting absence of annotation, while $R_{12}(i,j) = 1$ denoting the highest confidence of annotation. We define GO term $j^*$ as a candidate to annotate gene $i$ if the association score $R_{12}(i,j^*)$ is larger than the mean of association scores over all known annotations of gene $i$. To identify only high confidence gene-GO term predictions, we pick $j^*$ that are in the top 5% of largest association scores between GO term $j^*$ and all other genes. As before, we run our algorithm with and without GDV similarities (we exclude GDV similarities of Co-Ex network for reasons presented above).

We compute the percentage of reproduced, high confidence GO terms for CC, BP and MF separately. The results are shown in Figure 3a. Better results are achieved when GDV similarity matrices are included in the prediction model. Specifically, we capture 41% of BP terms, 41% of MF terms and 48% of CC terms. The BP and MF results outperform those of Dutkowski et al. (2013), whereas they achieve a higher percentage of reproduced GO terms in CC.

3.3 Validating predictions

Among all the statistically significant ($P \leq 0.01$) GO term association predictions, we find 132 not presented in GO (see Supplementary Table S1). To further increase confidence, we extract these associations from clusters with fewer than three GO terms that are stable over multiple factorization runs. We find that 14 of the 132 associations are between GO terms that have high semantic similarity and also confirm that additional 31 associations agree with predictions of Dutkowski et al. (2013). For example, our approach predicts term GO:003526 (Nua4 histone acetyltransferase complex) as a parent of GO:0043777 (Picollo Nua4 histone acetyltransferase complex), which was also reported by Dutkowski et al. (2013) and submitted to the GO Consortium for inclusion into the ontology. We further perform literature curation to validate the remaining predicted GO associations. We find literature support for another 13 of them (Supplementary Table S1). Hence, we validate 58 of 132 of our predictions.

Our approach not only identifies novel GO term association but it also makes highly reliable predictions for new gene-GO term relations. We predict new functional annotation of 972 genes (see Supplementary Table S2). Highly reliable predictions are those with association strength in the top 5%, as described in Section 3.2. For instance, we predict three genes, YDR101C, YDR49C and YNL132W, to be involved in ribosomal subunit biogenesis (GO:0042273) and find that the same functional prediction was previously reported through different approaches by Chen and Xu (2004) and Joshi et al. (2004). To validate the 972 predicted annotations, we use the new unpublished full set of yeast’s GI profiles from Boone Lab (Boone, 2014). The data consist of Pearson’s correlation coefficients of genetic profiles between gene pairs. We create the distribution of these correlations between newly annotated gene pairs for which we predict GO annotations. We compare this distribution of genetic profile correlations between the same number randomly sampled pairs of genes (we sampled multiple times and got consistent results). We observe higher correlations for predicted gene pairs than for random pairs (Fig. 3b). Moreover, using two-sample Kolmogorov–Smirnov (KS) test, we show that these two distributions significantly differ (KS statistics, $D = 0.2$ and $P$-value, $P = 1.5 \times 10^{-5}$). Thus, these results are highly consistent with our predictions of new annotations. This validates our predicted
GO annotations. Even though GI profiling analysis provides evidence that our algorithm is able to successfully predict new gene functions, additional biological validation would be needed for better understanding of these newly assigned functions.

4 CONCLUSIONS

We introduce a method for reconstruction of GO that is based on integrating solely the topology of biological networks. It captures 96% of the existing GO term relations and is capable of successfully identifying additional GO term associations as well as predicting gene annotations. Our method is general in the sense that it can integrate any heterogeneous systems-level interaction data. Therefore, it can easily be extended with new data that could consequently enhance the model’s predictive performance. This work suggests that the entire GO could be reconstructed from molecular interaction networks.

ACKNOWLEDGMENT

We thank Professor Charlie Boone for giving us his unpublished data for yeast.