Random walking on functional interaction networks to rank genes involved in cancer

Matteo Re and Giorgio Valentini

DI, Dipartimento di Informatica,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{re,valentini}@di.unimi.it

Abstract. A large scale analysis of gene expression data, performed by Segal and colleagues, identified sets of genes named Cancer Modules (CMs), involved in the onset and progression of cancer. By using functional interaction network data derived from different sources of biomolecular information, we show that random walks and label propagation algorithms are able to correctly rank genes with respect to CMs. In particular, the random walk with restart algorithm (RWR), by exploiting both the global topology of the functional interaction network, and local functional connections between genes relatively close to CM genes, achieves significantly better results than the other compared methods, suggesting that RWR could be applied to discover novel genes involved in the biological processes underlying tumoral diseases.

1 Introduction

The huge amount of data produced by large scale microarray experiments yielded to the development of specialized data repositories for effectively mining gene expression data related to cancer [1]. The availability of this unprecedented volume of data has, on the one hand, the potential to boost the research focused on the elucidation of the molecular basis of cancer and, on the other hand, to accelerate the development of novel cancer therapies. In this context, gene expression profiling proved to be effective for the discovery of subtypes of tumors [2], for the prediction of patients outcome [3] and the prediction of the response to chemotherapies [4].

In [5] expression profiles have been analyzed to identify sets of genes that act in concert to carry out specific functions in different cancer types, and to construct a collection of gene sets associated to specific Cancer gene Modules (CMs, hereafter). Nevertheless, even if gene expression data are fundamental to identify CMs, they cannot detect genes involved, for instance, in post-transcriptional, translational or post-translational misregulated processes underlying cancer. To take into account these post-transcriptional events, we analyzed integrated functional interaction networks obtained from curated databases and from protein-protein and protein domain-domain interactions, from protein complexes and from comparative genomics techniques [6, 7]. More precisely, we applied network

based algorithms to these functional interaction networks, in order to rank genes with respect to their likelihood to belong to specific CMs. By successfully ranking genes using functional interaction networks, we could in perspective discover novel genes involved in cancer, not detectable using data limited to correlated gene expression profiles.

Different algorithms to rank genes in biomolecular networks have been proposed in the literature [8, 9, 10, 11]. In this context networks are usually represent through an undirected graph $G = (V, E)$, where nodes $v \in V$ correspond to genes, and edges $e \in E$ are weighted according to the evidence of the functional interaction between genes [6, 7]. By exploiting proximity relationships between connected nodes, these algorithms are able to transfer annotations from previously annotated (labeled) nodes to unannotated (unlabeled) ones through a learning process inherently transductive in nature [12]. They include guilt-by-association methods and their extensions [9, 13], approaches based on the evaluation of the functional flow in graphs [14], methods based on Hopfield networks [15], and label propagation algorithms based on Markov [16] and Gaussian Random Fields [17].

Most of the cited approaches share the common feature of propagating known gene labels across the network, by exploiting the weighted connections between genes, until a certain criterion of convergence is satisfied. These approaches exploit the global topology of the network to rank genes, but in the context of functional interaction networks they could suffer the common drawback of exploring too far similarities between genes, thus introducing noise in the ranking process.

To deal with these problems in this work we apply random walk and random walk with restart algorithms [18] to rank genes with respect to their likelihood to belong to specific CMs. Indeed random walks can reduce the in-depth exploration of the network by limiting the number of allowed random steps, thus avoiding to consider too loose similarities between genes; moreover random walk algorithms have been recently successfully applied to the related problem of gene prioritization of candidate disease genes [19]. We compared random walks algorithms with other state-of-the-art gene ranking algorithms through an extensive experimental analysis involving about 300 CMs and functional interaction networks with more than 8000 human genes and hundreds of thousands of edges connecting genes.

2 Methods

In this section we introduce *random walk* (*RW*), *random walk with restart* (*RWR*) and three variants of *label propagation* algorithms used in our experiments to rank genes with respect to CMs.

All the methods described below refer to an undirected weighted graph $G = (V, E)$, where nodes $i, j \in V$ correspond to genes, with $|V| = n$, and edges $(i, j) \in E$ are weighted according to the weight matrix $\boldsymbol{W}$, whose elements w_{ij}

are the weights of the edges (i, j), and represent the"strength" of the functional interaction between genes i and j.

2.1 Random walk and random walk with restart

A *random walk* (*RW*) on G is a reversible Markov chain with transition matrix $\boldsymbol{Q}$, whose elements q_{ij} satisfy the probabilistic constraint $\sum_j q_{ij} = 1$:

$$q_{ij} = w_{ij} / \sum_k w_{ik} \tag{1}$$

In the context of gene ranking with respect to CMs, *RW* algorithms [18] explore and exploit the topology of the functional network, starting and walking around from a subset $V_M \subset V$ of genes belonging to a specific Cancer Module M by using a transition probability matrix $\boldsymbol{Q} = \boldsymbol{D}^{-1}\boldsymbol{W}$, where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $d_{ii} = \sum_j w_{ij}$. The elements q_{ij} of $\boldsymbol{Q}$ represent the probability of a random step from i to j. The initial probability of belonging to M can be set to $p^o = 1/|V_M|$ for the genes $i \in V_M$ and to $p^o = 0$ for the genes $i \in V \setminus V_M$. If $\boldsymbol{p}^t$ represents the probability vector of finding a "random walker" at step t in the nodes $i \in V$ (that is, p_i^t represent the probability for a random walk of reaching node i at step t), then the probability at step $t + 1$ is:

$$\boldsymbol{p}^{t+1} = \boldsymbol{Q}^T \boldsymbol{p}^t \tag{2}$$

and the update (2) is iterated until convergence. We can observe that in the context of gene functional interaction networks, by running the algorithm in an iterative way until, for a given t, $\boldsymbol{p}^t = \boldsymbol{p}^{t+1}$, we could progressively "forget" the a priori information available for the Cancer Module M: in other words we could explore nodes too far from the "core" nodes included in V_M, thus introducing functional similarities between genes even when no functional interactions are actually present between genes. To avoid these drawbacks, we can stop the random walk before convergence, thus considering only meaningful functional relationships between genes.

This approach requires to experimentally find the "optimal" number of random steps, or simply to try with different number of predefined steps. An alternative approach is represented by the *random walk with restart* (*RWR*) algorithm [18]: at each step the random walker can move to one of its neighbours or can restart from its initial condition with probability θ:

$$\boldsymbol{p}^{t+1} = (1 - \theta)\, \boldsymbol{Q}^T \boldsymbol{p}^t + \theta\, \boldsymbol{p}^o \tag{3}$$

Fig. 1 shows the pseudocode of the *RWR* algorithm. With both *RW* and *RWR* methods at the steady state we can rank the vector $\boldsymbol{p}$ to prioritize genes according to their likelihood to belong to the CM under study.

Fig. 1. Random walk wih restart algorithm

`Input`:
- $\boldsymbol{W}$: weight matrix of the graph
- $V_M \subset V$: genes belonging to a cancer module M
- ϵ: convergence parameter
- θ: restart probability

`begin algorithm`
01: `for each` $i \in V_M$ $p_i^o := 1/V_M$
02: `for each` $i \notin V_M$ $p_i^o := 0$
03: `for each` $i \in V$ $d_{ii} := \sum_j w_{ij}$
04: $\boldsymbol{Q} := \boldsymbol{D}^{-1}\boldsymbol{W}$
05: t:=0
06: `repeat`
07: t := t + 1
07: $\boldsymbol{p}^t = (1-\theta)\,\boldsymbol{Q}^T\boldsymbol{p}^{t-1} + \theta\,\boldsymbol{p}^o$
08: `until` $(||\boldsymbol{p}^t - \boldsymbol{p}^{t-1}|| < \epsilon)$
09: `for each` $i \in V$
10: $p_i^t := p_i^t / \sum_j p_j^t$
`end algorithm`.
`Output`: the probability vector $\boldsymbol{p}^t$

2.2 Label propagation algorithms

These algorithms are characterized by the propagation of the information from a "core" of labeled nodes to an usually larger set of the unlabeled nodes of the graph under study, by a semi-supervised transductive learning process [12]. From this standpoint, this approach resembles random walks, and the process is iterated until to convergence by minimizing a quadratic objective function. For instance, the label propagation algorithm *LP* proposed in [20] minimizes the following objective function

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{s}} \sum_i \sum_j w_{ij}(s_i - s_j)^2 \tag{4}$$

where $\boldsymbol{s}$ is the score vector used to rank the genes (similar to the probability vector $\boldsymbol{p}$ of the random walks), and s_i is its i^{th} component relative to the i^{th} gene. Eq. (4) represents the "internal coherence" of the network: it penalizes connected genes (i.e. pairs of genes i and j with $w_{ij} > 0$) having different scores. *LP* assures the coherence with respect to the initial score $\boldsymbol{s}^0$ by not allowing any change of the scores s_i for the vertices $i \in V_M$ during the label propagation process: the predicted scores s_i are set to s_i^0 for each $i \in V_M$. This algorithm can be implemented through iterative techniques or in closed form by solving a system of linear equations.

By minimizing a quadratic objective function that directly embeds a "fitting term" representing the error between predicted and a priori known scores, we

can derive a regularized label propagation algorithm *LPR* proposed in [21]:

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{s}} \alpha \sum_i (s_i - s_i^0)^2 + (1-\alpha) \sum_i \sum_j w_{ij}(s_i - s_j)^2 \qquad (5)$$

where $\boldsymbol{s}$ is the vector of the scores associated to the genes, $\boldsymbol{s}^0$ is the initial vector of scores reflecting the a priori knowledge about the investigated genes, s_i and s_i^0 their i^{th} components, and w_{ij} are the elements of the weight matrix $\boldsymbol{W}$ of the graph G connecting the genes. Note that (5) is the convex combination ($0 \leq \alpha \leq 1$) of two terms, where the first one minimizes the error between predicted and a priori known scores (the "fitting term"), while the second assures the "internal coherence" of the network, and is analogous to (4).

A variant of *LPR* algorithm is represented by *GeneMANIA* [22], originally proposed to predict gene functions. This algorithm introduces a simple but effective cost-sensitive technique (useful when the number of positive examples is largely lower than the total number examples), and moreover minimizes (5) through an efficient iterative algorithms based on conjugate gradient techniques.

3 Cancer Module genes ranking

In this section, at first we describe the Cancer gene Modules (CMs) proposed in [5] and the functional interaction networks used in our experiments to rank genes according to their likelihood to belong to specific CMs. Then we present the experimental set-up and discuss the results obtained with the gene ranking network-based methods introduced in Section 2.

3.1 Functional interaction networks and Cancer Modules

Segal and colleagues analyzed at genome-wide level the human gene expression profiles of about 2000 arrays spanning 17 clinical categories represented by several types of tumor. By considering about 3000 publicly available gene sets they identified 456 statistically significant gene sets called Cancer Modules (CMs) by the authors (see [5] for further details). The authors mapped coordinately over or underexpressed modules to the clinical conditions associated to each array, thus characterizing different types of tumor in terms of sets of altered functional gene modules.

To rank genes with respect to the CMs we considered two types of functional interaction networks: the first one is a functional protein interaction network (*FI*) based on interactions provided by a Naive-Bayes classifier [6]; the second is a functional human gene network (*HumanNet*) that has been used in several tests to predict causal genes for human diseases and to increase the power of genome-wide association studies [7]. More precisely *FI* is based on functional interactions predicted by a Naive Bayes classifier (NBC) trained on pairwise relationships extracted from Reactome[23] and other curated pathways databases, and from uncurated pairwise relationships derived from physical protein-protein

interactions (PPI) in human and other species, from gene co-expression data, proteins domain-domain interactions, protein interactions obtained via biomedical text mining, and Gene Ontology annotations. *HumanNet* is characterized by functional interactions derived from different species through comparative genomics techniques, by which functional interactions are propagated from different model organisms to human by means of a comparative genomics approach presented in [24].

3.2 Experimental set-up

To avoid singleton nodes in the functional networks (Sect. 3.1), we removed genes with no functional interactions with any other gene in *FI* or *HumanNet* networks. Moreover, to assure reasonably reliable predictions, we removed CMs annotated with less than 20 genes, thus resulting in 298 CMs and a collection of about 8500 human genes (nodes of the networks). Genes were ranked with respect to each CM: for each of the 298 CMs we computed both the precision at fixed recall rates and the area under the ROC curve (AUC), by adopting a 5-fold stratified cross-validation (CV) technique. We compared the average results across CMs obtained with random walks at 1, 2 and 3 steps, random walks with restart, and with the label propagation algorithms summarized in Section 2.2. In our experiments we set $\theta = 0.6$ for *RWR* and $\alpha = 0.7$ for *LPR* algorithms.

3.3 Results

Fig. 2 show the compared precision at recall rates varying from 0.1 to 1 by 0.1 steps obtained by the different methods. Note that the results are averaged across the 298 CMs. At any recall rate *RWR* significantly outperforms any other compared method with both the *FI* (Fig. 2, top) and *HumanNet* (Fig. 2, bottom) networks. Note that at recall rates from 0.1 to 0.3 *RWR* achieves a precision equal to 1 or very close to 1 with both *FI* and *HumanNet* networks, showing that top ranked genes are all true positive genes for all the the considered CMs, with no false positives. The second best method in term of precision/recall is *LP*: indeed with *FI* and especially with *HumanNet* for recall levels from 0.1 to 0.5 obtains the best results, but for higher recall rates 2-steps *RW* achieves slightly better results than *LP*. The worst performances have been obtained by 1 and 3-steps *RW*, while the other label propagation algorithms (*GeneMANIA* and *LPR*) registered intermediate results between 2-steps *RW* and 1 and 3-steps *RW* (Fig. 2). All the methods monotonically decrease their performance moving toward high recall rates. The only exception is due to *LP* that at first increases and then decreases precision rates (Fig. 2); this means that at low recall rates (that is when we consider the top ranked positive genes) *LP* has a relatively large rate of false positives (lower precision), but it achieves a better precision at higher recall rates, that is *LP* better ranks positive genes that are just below the top ranked genes.

These general trends are confirmed also by AUC results averaged across the CMs (Fig. 3). *RWR* significantly outperforms all the other methods at 10^{-5}

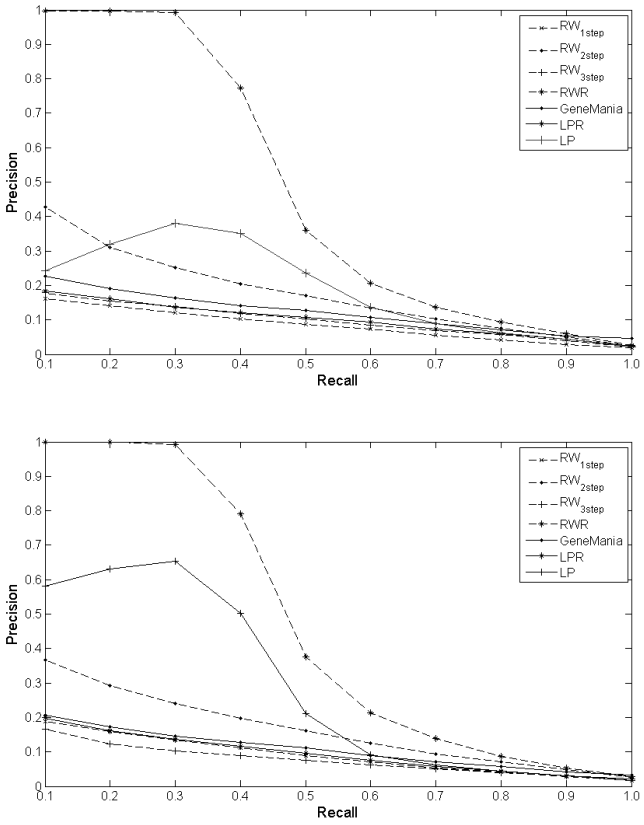

Fig. 2. Compared precision at a given recall level results. Top: *FI* network; Bottom: *HumanNet* network. *RW* stands for random walk, *RWR* for random walk with restart, *LP* for label propagation and *LPR* for label propagation regularized algorithms.

significance level, according to the Wilcoxon rank sum test, with both *FI* and *HumanNet* networks. Quite interestingly, 2-steps *RW* is the second best method, significantly better also than *LP* (Wilcoxon rank sum test, 10^{-5} significance level): even if *LP* shows a better precision at low recall rates (Fig. 2), on the average 2-steps *RW* ranks better positive genes (Fig. 3). The other methods behave significantly worse, with *GeneMANIA* slightly better than its non cost-sensitive counterpart *LPR* (CMs are quite unbalanced with on the average a significant lower number of positive genes), and the worst method is 1-step *RW* (Fig. 3).

These results altogether show that successful methods in this complex ranking task are those able to exploit the functional relationships connecting genes relatively close to the "core" of positive genes V_M, but at the same time able to

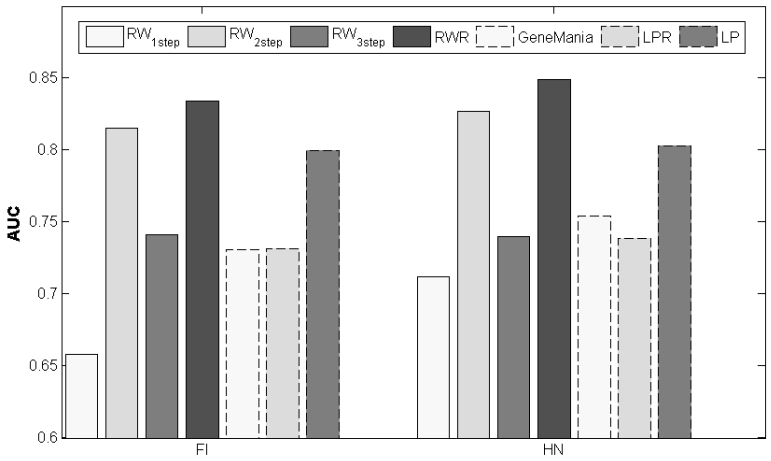

Fig. 3. Compared AUC results. On the left side are grouped results relative to the *FI* network, on the right results relative to the *HumanNet* network. *RW* stands for random walk, *RWR* for random walk with restart, *LP* for label propagation and *LPR* for label propagation regularized algorithms.

learn from the global topological characteristics of the functional networks. Indeed, considering the *RW* algorithm, 2-steps *RW* achieves the best results, while both 1-step and 3-steps *RW* behaves worse on this ranking task (and 4-steps also worse than 3-steps *RW*, data not shown). Moreover if we run the *RW* algorithm until to convergence, we achieve average AUC close to 0.5 (that is no learning). This means that only functional relationships connecting genes relatively close to the set V_M of positive genes are useful to learn the CM. This fact is confirmed also by the results obtained with the *LP* label propagation algorithm; indeed we stopped it after 20 iterations, while the original algorithm [20] stops only at convergence, and if we follow the original algorithm we obtain also in this case an average AUC close to 0.5 (data not shown). This means that exploring genes (nodes) too far from the set V_M of positive genes, we may add noise to the label ranking procedure: genes are considered similar even when paths are too long to preserve a significant functional similarity between them. However, for several CMs we obtained the best results with 2, 3 or in some cases also more steps (data not shown), thus suggesting that for at least some CMs the overall topology of the network is useful to rank genes. The *RWR* algorithm on the one hand takes into account the nodes/genes close to the V_M set of positive genes through the "restart" mechanism, but on the other hand exploits also the overall topology of the networks, by walking across the network until to convergence (Section 2.1). We think that these features of the *RWR* algorithm fit well the characteristics

of the functional networks and can explain the good results obtained by this method.

4 Conclusions

CMs were defined mainly with expression signatures obtained from gene expression data profiling [5]. We show that network-based algorithms can successfully rank genes with respect to CMs, by using functional interaction networks constructed from physical protein-protein interactions, proteins domain-domain interactions, and other sources of biomolecular information.

In particular *random walk with restart* algorithms significantly outperform all the other compared methods based on limited steps random walks and state-of-the-art label propagation algorithms, showing that we need learning algorithms able to learn form both the global topology of the functional network and the functional relationships closer to the set of positive genes. The very high precision achieved at different levels of recall by the *RWR* algorithm with about 300 CMs using two functional networks including more than 8000 human genes, suggests that this method could be in perspective applied to discover novel genes involved in the onset and progression of cancer.

Acknowledgments

We would like to thank the reviewers for their comments and suggestions. The authors gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

References

[1] Rhodes, D., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A.: Oncomine: a cancer microarray database and integrated data-mining platform. Neoplasia **6** (2004) 1–6

[2] Alizadeh, A., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511

[3] van't Veer, L., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(31) (2002) 530–536

[4] Straver, M., Glas, A., Hannemann, J., Wesseling, J., van de Vijver, M., Rutgers, E., Vrancken Peeters, M., van Tinteren, H., Van't, L., S, R.: The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. Breast Cancer Res Treat **119** (2010) 551–558

[5] Segal, E., Friedman, N., Koller, D., Regev, A.: A module map showing conditional activity of expression modules in cancer. Nat Genet **36** (2004) 1090–1098

[6] Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Genome Biology **11** (2010) R53

[7] Lee, I., Blom, U., Wang, P., Shim, J., Marcotte, E.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. **21** (2011) 1109–1121
[8] Aerts, S., et al.: Gene prioritization through genomic data fusion. Nature Biotechnology **24**(5) (2006) 537–544
[9] McDermott, J.and Bumgarner, R., Samudrala, R.: Functional annotation from predicted protein interaction networks. Bioinformatics **21**(15) (2005) 3217–3226
[10] Erten, S., Bebek, G., Koyuturk, M.: Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. J Comput Biol. **18**(11) (2011) 1561–1574
[11] Re, M., Valentini, G.: Cancer module genes ranking using kernelized score functions. BMC Bioinformatics (2012) (in press).
[12] Bengio, Y., Delalleau, O., Le Roux, N.: Label Propagation and Quadratic Criterion. In: Semi-Supervised Learning. MIT Press (2006) 193–216
[13] Re, M., Valentini, G.: Large scale ranking and repositioning of drugs with respect to DrugBank therapeutic categories. In Bleris, L., et al., eds.: International Symposium on Bioinformatics Research and Applications (ISBRA 2012). Lecture Notes in Computer Science, Springer (2012) 225–236
[14] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21**(S1) (2005) 302–310
[15] Bertoni, A., Frasca, M., Valentini, G.: Cosnet: a cost sensitive neural network for semi-supervised learning in graphs. In: European Conference on Machine Learning, ECML PKDD 2011. Volume 6911 of Lecture Notes on Artificial Intelligence., Springer (2011) 219–234
[16] Deng, M., Chen, T., Sun, F.: An integrated probabilistic model for functional prediction of proteins. J. Comput. Biol. **11** (2004) 463–475
[17] Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q.: GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology **9**(S4) (2008)
[18] Lovasz, L.: Random Walks on Graphs: a Survey. Combinatorics, Paul Erdos is Eighty **2** (1993) 1–46
[19] Li, Y., Patra, J.: Integration of multiple data sources to prioritize candidate genes using discounted rating systems. BMC Bioinformatics **11**(Suppl I:S20) (2010)
[20] Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In: Proc. of the Twentieth International Conference on Machine Learning, Washington DC (2003) 912–919
[21] Zhou, D., Bousquet, O., Lal, T., Weston, J., Scholkopf, B.: Learning with Local and Global Consistency. In: Advances in Neural Information Processing Systems. Volume 16., Cambridge, MA, MIT Press (2004) 321–328
[22] Mostafavi, S., Morris, Q.: Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics **26**(14) (2010) 1759–1765
[23] Vastrik, I. et al: Reactome: a knowledge base of biologic pathways and processes. Genome Biol. **8** (2007) R39
[24] Lee, I., Li, Z., Marcotte, E.: An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. PLoS ONE **2** (2007) e988