Disease–Genes must Guide Data Source Integration in the Gene Prioritization Process

Marco Frasca[(1)], Jean Fred Fontaine[(2)], Giorgio Valentini[(1)], Marco Mesiti[(1)], Marco Notaro[(1)], Dario Malchiodi[(1)] and Miguel A. Andrade-Navarro[(2)]

(1) Anacleto Lab – Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39, 20135. frasca@di.unimi.it, malchiodi@di.unimi.it

(2) Faculty of Biology, Johannes Gutenberg University Mainz and Institute of Molecular Biology, Ackermannweg 4, Mainz 55128, Germany. Andrade@uni-mainz.de, fontaine@uni-mainz.de

Keywords: Medical Subject Headings, gene prioritization, imbalance-aware integration, network integration.

Abstract. One of the main issues in the problem of detecting genes involved in the etiology of genetic human diseases is the integration of different types of available functional relationships between genes. Numerous approaches exploited the complementary evidence coded in heterogeneous sources of data to prioritize disease-genes, such as functional profiles or expression quantitative trait loci, but none of them to our knowledge posed the scarcity of known disease-genes as a feature of their integration methodology. Nevertheless, in contexts where data are unbalanced, that is, where one class is largely under-represented, imbalance-unaware approaches may suffer a strong decrease in performance. We claim that imbalance-aware integration is a key requirement for boosting performance of gene prioritization (GP) methods. To support our claim, we propose an imbalance-aware integration algorithm for the GP problem, and we compare it on benchmark data with other integration methodologies neglecting data imbalance.

1 Scientific Background

In the context of Network Medicine, discovering genes causing complex diseases, also known as "disease-genes", has become a central and complex challenge [1, 2]. This process, called gene prioritization (GP), usually aims to supply a ranking of genes according to their involvement in the etiology of a given disease. A main issue characterizing the GP problem is the availability of a large amount of heterogeneous information about gene networks, ranging from protein–protein interactions to gene co-expression and functional similarity [3]. Indeed, excluding the potentially complementary evidence coming from heterogeneous data sources may be a strong limitation [4]. Several research groups have adopted computational methodologies that rely on the use of multiple heterogeneous networked-sources, and a general approach is to combine the topology of each available network into a more informative 'consensus' network, also having a larger coverage [5, 6]. A common practice leverages weighted schemas to construct a linear combination of the input networks, by computing for the disease under study an informativeness coefficient for each network. This coefficient represents, to some extent, the 'usefulness' of the network in predicting that disease, and it should take into account the rarity of known disease-genes characterizing most diseases in existing disease ontologies, such as the Medical Subject Headings (MeSH)[1] (thousands of genetic diseases still have none or very few known causative genes). When a disease-gene (positive gene) is rare for a given disease, it carries most information about the

[1] http://www.nlm.nih.gov/mesh/

latter, and when an input source is such that it helps classifiers in correctly ranking positive genes, the informativeness coefficient of that source must be higher. Indeed, imbalance-aware integration led to successful results in similar contexts, such as the protein function prediction [7]. Nevertheless, this central issue has been neglected by most existing approaches for data source integration for gene prioritization.

We argue in this paper that, even for the GP problem, network integration must consider the rarity of available disease-genes to improve the accuracy of gene rankings. To this purpose, we extended the integration algorithm *UNIPred* (Unbalance-aware Network Integration and Prediction, [7]), in order to emphasize the importance of positive genes in the integration process. UNIPred can capture the usefulness of each data source for a disease of interest by specifically handling the imbalance of data labelings. The method has been extended by introducing a novel optimization criterion, in which the relevance to be attributed to positive genes is associated with a free parameter, so as to facilitate verifying our claim. By using the network usefulness computed through UNIPred, the consensus network is built and given as input to *WGP*, a recent network-based algorithm proposed to prioritize disease-genes [8]. The overall methodology has been then validated on a benchmark data set composed of nine human networks and 708 MeSH disease terms [5].

2 Materials

Our setup follows a benchmark proposed in [5] for data integration in the GP context. Nine human gene networks covering 8449 genes are available, considering heterogeneous data sources, as described in the following (see [5] for details about each network).

Functional interaction network – *finet.* A network covering 8441 selected proteins and contains protein–protein functional binary interactions.

Human net – *hnnet.* A network combining distinct lines of evidence from four species, including human mRNA co-expression, protein–protein interactions, and protein complexes.

Cancer module network – *cmnet.* A network of 8849 genes collecting interactions derived from expression profiles in different tumors in terms of the behavior of modules of correlated genes.

Gene chemical network – *gcnet*. A network of 7649 genes constructed on the basis of direct and indirect genes–chemical interactions available at the Comparative Toxicogenomics Database (CTD) [9].

BioGRID database network – *dbnet*. Protein–protein interaction network for 8449 proteins based upon direct physical and genetic interactions obtained from BioGRID (v. 3.2.96 - January 2013).

BioGRID projected network – *bgnet*. An extended network from BioGRID constructed by retrieving the connection between the 8849 genes in the benchmark against all human genes in a bipartite graph, and by considering the common neighbours to determine the degree of similarity between two genes in the benchmark.

Semantic similarity networks – {*bp,mf,cc*}*net*. Three networks obtained by considering the Gene Ontology (GO, [10]) terms in the three branches annotating the considered genes: biological process (bp), molecular function (mf) and cellular component (cc).

Gene–disease associations have been downloaded from the CTD database and include 708 selected MeSH terms having from 5 to 200 annotated disease-genes.

3 Methods

Every input network is represented through a weighted undirected graph $G^{(k)} = \langle V, \boldsymbol{W}^{(k)} \rangle$ on the genes/instances V (or a subset of it), where $k \in \{1, 2, \ldots, m\}$ is the network index and $\boldsymbol{W}^{(k)}$ is the connection matrix: the entry $W^{(k)}_{i,j} \in [0, 1]$ indicates a degree of functional similarity between genes i and j. Here m is the number of available networks over genes V. If a network covers just a subset of V, it is extended to V by adding zeros in the corresponding entries of the connection matrix. We assume thereby in the following that all networks cover the set V. Given a disease of interest d, every gene $i \in V$ possesses a label $y_i \in \{0,\ 1\}$ denoting that gene i is currently associated with d (label 1, positive gene) or not (label 0, negative gene).

The aim is to construct a composite network $G = \langle V, \boldsymbol{W} \rangle$ integrating all available networks. This is performed by associating every network $G^{(k)}$ with a coefficient $r^{(k)}_d$ related to its informativeness for disease d. To compute $r^{(k)}_d$ we adopt an extension of the UNIPred algorithm, briefly described in the following.

3.1 UNIPred

The UNIPred algorithm computes for every networked-source a relevance score taking expressly into account the disproportion between 1-labeled and 0-labeled genes for the studied genetic disease d. In particular, UNIPred operates a network projection onto the plane so that each gene $i \in V$ is associated with a labelled bi-dimensional point $P^{(k)}_i$, embedding the local imbalance in the corresponding node position. For a given network $G^{(k)}$, the coordinates $P^{(k)}_i \equiv (P^{(k)}_{i,1};\ P^{(k)}_{i,2})$ are computed as follows:

$$\begin{aligned} P^{(k)}_{i,1} &= \sum_{j \in V} W^{(k)}_{ij} \cdot y_j \\ P^{(k)}_{i,2} &= \sum_{j \in V} W^{(k)}_{ij} \cdot (1 - y_j) \end{aligned} \tag{1}$$

In other words, $P^{(k)}_{i,1}$ is the weighted sum of 1-labeled neighbors, $P^{(k)}_{i,2}$ is the weighted sum of 0-labeled neighbors. The position of each point in the plane thereby reflects the topology of the connections towards neighboring positive and negative nodes.

The algorithm then learns a parametric straight line to separate positive and negative points. In this way, since every point already has a label, each line is associated with the number $TP^{(k)}$ of positive points correctly classified (true positives), the number $FN^{(k)}$ of positive points wrongly classified (false negatives), and the number $FP^{(k)}$ of negative points wrongly classified (false positives). Lines are selected by an approximated supervised algorithm maximizing an imbalance-aware criterion, namely the ***F–measure***: $F^{(k)} = \frac{2TP^{(k)}}{2TP^{(k)} + FP^{(k)} + FN^{(k)}}$. The maximum value $\bar{F}^{(k)}$ of the F–measure obtained during the learning phase is then considered as relevance $r^{(k)}_d$ for the input network $G^{(k)}$. In order to emphasize the need of attributing higher importance to positive genes, here we introduce a variant of the objective function, named F_β***–measure***, defined as:

$$F^{(k)}_\beta := \frac{(1 + \beta^2) TP^{(k)}}{(1 + \beta^2) TP^{(k)} + FP^{(k)} + \beta^2 FN^{(k)}} \tag{2}$$

Indeed, the parameter $\beta \in \mathbb{R}^+$ allows to regulate the importance to be assigned to the misclassification of positives rather than negatives, thus for $\beta > 1$ we penalize more the misclassification of positives. The larger β, the more relevant are positives in determining the network coefficient $r^{(k)}_d$.

3.2 Constructing the integrated network

For a given disease of interest d, UNIPred is applied to each input network independently, obtaining the relevance vector $\boldsymbol{r}_d = (r_d^{(1)},\ r_d^{(2)}, \cdots,\ r_d^{(m)})$. The consensus network is then constructed as a weighted sum of the corresponding adjacency matrices:

$$\boldsymbol{W} = \sum_{k=1}^{m} r_d^{(k)} \boldsymbol{W}^{(k)}$$

Moreover, in order to have a baseline comparison, networks are also integrated by unweighted average sum (*US*) that is $\boldsymbol{W} = \sum_{k=1}^{m} \boldsymbol{W}^{(k)} / m$.

3.3 Inferring the gene prioritization list

Once the consensus network $G = \langle V, \boldsymbol{W} \rangle$ is constructed, we are ready to face the gene prioritization problem, which is modeled as a semi-supervised ranking problem on graphs. The set of genes is assumed to be partitioned into L and U, disjoints subsets of V respectively containing the labeled and unlabeled genes, and the objective is to infer a ranking of genes in U according to the given disease of interest. Only for genes $i \in L$ the label $y_i \in \{0,\ 1\}$ is thereby known (with reference to the studied disease d), and the aim is learning a function $\phi : U \to \mathbb{R}$ so as to rank higher genes susceptible to be involved in the etiology of d.

Furthermore, analogously to the integration step, the complexity of the problem is increased by the rare presence of positive genes, necessary for any classifier to infer 'meaningful' solutions. Accordingly, the adopted methodology has to consider this characteristic of the problem to prevent a large decay of the ranking quality [11]. To learn the ranking function ϕ we employed a regression model proposed in [8], termed *WGP* (*Weighted Gene Prioritization*), able in handling the label imbalance. Briefly, starting from the integrated network, WGP adopts a node projection onto a bi-dimensional space, where a weighted binomial regression model with $\log$-$\log$ link function, a skewed function suitable for unbalanced data, is learned to separate positive and negative classes, and to consequently infer the desired prediction.

4 **Results**

Following the benchmark setting [5], the generalization performance of our method has been assessed through a classical 5-fold cross-validation procedure, and the results were evaluated by using the Area Under the Receiver Operating Characteristic Curve (*AUC*) and the Precision at different Recall levels (*PxR*). In addition, we computed the Area Under the Precision Recall Curve (*AUPRC*), to take into account the imbalance of annotated vs. unannotated genes for the MeSH disease terms. The validation of WGP-UNIPred algorithm on benchmark data showed a noticeable improvement with respect to the compared methods, including random walks [12], random walks with restarts, guilt-by-association methods [13] and kernelized average score functions (S_{AV}, [14]). Fig. 1 shows the overall performance, emphasizing both the gain of UNIPred with respect to US integration schema and the influence of the β parameter on the performance. In Fig. 1 and 2 we only showed the results of S_{AV} with weighted (WS) and unweighted sum (US) integration, since random walk and the other compared methods achieved worse results than S_{AV}. In particular, in [5] average AUC results across diseases have been used to weight networks according to the WS integration for S_{AV}.

To better evaluate the behaviour of our methodology, we also show results averaged across diseases with at most 10 (category 'l10') and more than 10 (category 'm10') associated genes. AUPRC results are not provided in the benchmark. The predictive capability of the model remarkably improves when increasing the parameter β, and more in the most unbalanced diseases (*l10*), confirming the need of imbalance-aware integration. Conversely, in US schemas, there is an almost negligible difference between

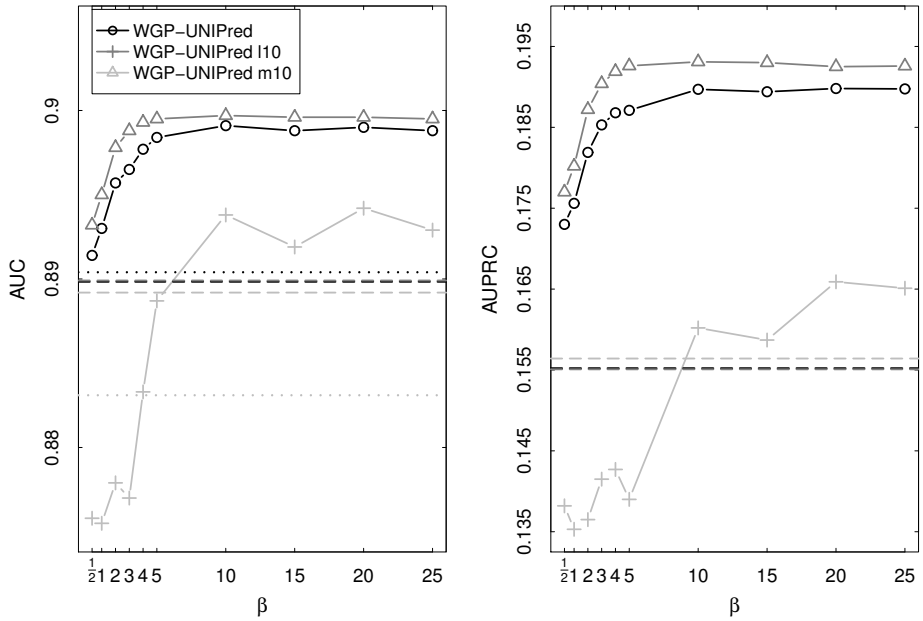

Figure 1: Performance of WGP-UNIPRED on benchmark data. 'l10' and 'm10' refer to the subsets of MeSH disease terms with 5–10 and 11–200 associated genes, respectively, whereas circles correspond to results averaged across all diseases. The black horizontal dashed line corresponds to the average performance across all diseases of WGP on unweighted sum (US) data, whereas the dashed light grey (resp. grey) line corresponds to WGP performance on US data averaged across the category 'l10' (resp. 'm10'). The black and light grey horizontal dotted lines for the AUC results instead denote the average AUC achieved by S_{AV} on weighted (WS) and unweighted sum data, respectively.

l10 and *m10* disease categories. The performance of WGP-UNIPred tends to become stable for values of β larger than 10, and interestingly, the improvement of weighted integration is larger for WGP than for S_{AV} when compared with the corresponding unweighted strategies. This confirms that using an imbalance-aware criterion (unlike the AUC) to weight networks is more effective in this context. Apparently, the larger improvement for UNIPred compared to US schema for *m10* with respect to *l10* terms (in both AUC and AUPRC) is quite unexpected, since *l10* terms are more unbalanced; nevertheless, since the available information for *l10* terms is very small, this behavior is likely due to overfitting phenomena. Indeed, similar works have shown that regularizing the network effectiveness for more unbalanced terms led to better results [15].

WGP-UNIPred ($\beta = 20$) also compares favourably in terms of PxR (Fig. 2). Even in this setting, WGP-UNIPred outperforms S_{AV}, except for 0.1 recall level on US data. Confirming the behaviour in terms of AUC, the weighted sum integration (WS) using the UNIPred imbalance-aware network relevance led to larger improvements than the imbalance-unaware weighted integration with regard to US corresponding results.

5 Conclusion

Experimental results supported our claim that the integration of omics data (genomics, transcriptomics, proteomics and so on) need imbalance-aware procedures for improving the accuracy of gene prioritization lists. A state-of-the-art integration algorithm, UNIPred [7], has been used to boost the performance of a gene prioritization method, WGP [8]. By explicitly modelling the integration procedure on the exploitation of the known disease-genes, WGP-UNIPred outperformed other state-of-the-art methods in predicting gene–disease associations on public benchmark data.

Acknowledgments

This work was partially funded by the Department of Computer Science of the University of Milan under the grant "Piano di sostegno alla ricerca 2015/2017" (grant num-

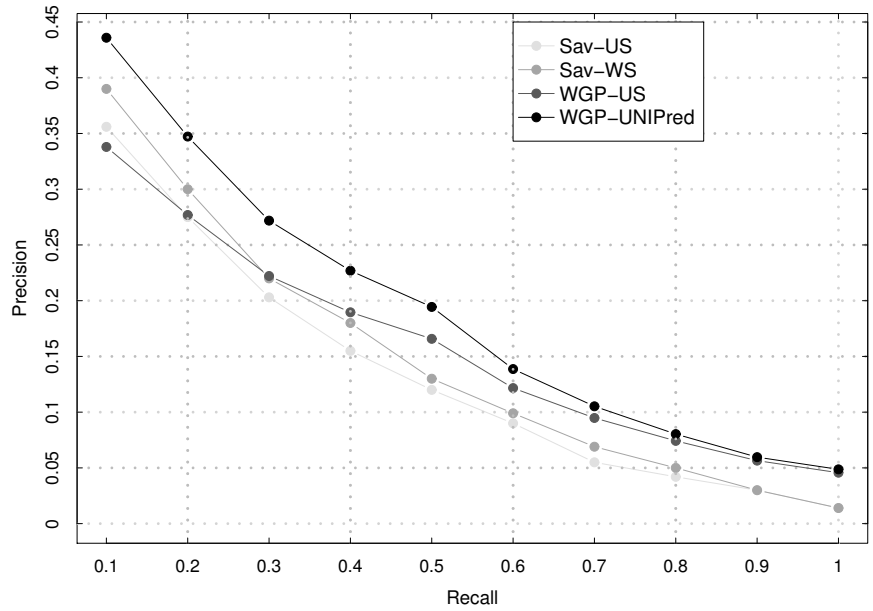

Figure 2: PxR results achieved by the top benchmark method S_{AV} and WGP-UNIPred on both unweighted and weighted schemas.

ber 19246), project title *Graph-based methodologies for the automated inference in bio-medical ontologies*.

References

[1] A. L. Barabasi, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, pp. 56–68, Jan. 2011.

[2] N. Tiffin, M. A. Andrade-Navarro, and C. Perez-Iratxeta, "Linking genes to diseases: it's all in the data," *Genome Medicine*, vol. 1, no. 8, p. 77, 2009.

[3] R. M. Piro and F. Di Cunto, "Computational approaches to disease-gene prediction: rationale, classification and successes," *FEBS Journal*, vol. 279, no. 5, pp. 678–696, 2012.

[4] J. Che and M. Shin, "A Meta-Analysis Strategy for Gene Prioritization Using Gene Expression, SNP Genotype, and eQTL Data," *BioMed Research International*, vol. 2015, pp. 1–8, 2015.

[5] G. Valentini *et al.*, "An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods," *Artificial Intelligence in Medicine*, vol. 61, no. 2, pp. 63 – 78, 2014.

[6] M. Frasca and D. Malchiodi, "Exploiting Negative Sample Selection for Prioritizing Candidate Disease Genes," *Genomics and Computational Biology*, vol. 3, no. 3, p. e47, 2017.

[7] M. Frasca, A. Bertoni, and G. Valentini, "UNIPred: Unbalance-aware Network Integration and Prediction of Protein Functions," *Journal of Computational Biology*, vol. 22, no. 12, pp. 1057–1074, 2015.

[8] M. Frasca and S. Bassis, "Gene-Disease Prioritization Through Cost-Sensitive Graph-Based Methodologies," in *IWBBIO 2016*, vol. 9656 of *Lecture Notes in Computer Science*, pp. 739–751, Springer, 2016.

[9] A. P. Davis *et al.*, "Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic acids research*, vol. 37, no. Database issue, pp. D786–92, 2009.

[10] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature genetics*, vol. 25, pp. 25–29, May 2000.

[11] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

[12] L. Lovász, "Random walks on graphs: A survey," in *Combinatorics, Paul Erdős is Eighty* (D. Miklós, V. T. Sós, and T. Szőnyi, eds.), vol. 2, pp. 353–398, Budapest: János Bolyai Mathematical Society, 1996.

[13] E. Marcotte, M. Pellegrini, M. Thompson, T. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function," *Nature*, vol. 402, pp. 83–86, 1999.

[14] G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti, and M. Re, "RANKS: a flexible tool for node label ranking and classification in biological networks," *Bioinformatics*, 2016.

[15] S. Mostafavi and Q. Morris, "Fast integration of heterogeneous data sources for predicting gene function with limited annotation," *Bioinformatics*, vol. 26, pp. 1759–1765, July 2010.