









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

Marco Frasca¹(✉) , Jean Fred Fontaine² , Giorgio Valentini¹ ,
Marco Mesiti¹ , Marco Notaro¹ , Dario Malchiodi¹ ,
and Miguel A. Andrade-Navarro²

¹ Anacleto Lab – Dipartimento di Informatica, Università degli Studi di Milano,
Via Celoria 18, 20133 Milan, Italy

{frasca,valentini,mesiti,malchiodi}@di.unimi.it, marco.notaro@unimi.it

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

Abstract. One of the main issues in detecting the 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 state-of-the-art integration methodologies.

Keywords: Medical Subject Headings · Gene prioritization · Imbalance-aware integration · Network integration

1 Background

In the context of Network Medicine, discovering genes causing or associated with complex diseases, also known as “disease-genes”, has become a central and complex challenge [2, 7, 16]. 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 genes, ranging

from protein–protein interactions to gene co-expression and functional similarity [15]. Excluding the potentially complementary evidence coming from heterogeneous data sources may be a strong limitation [3]. 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 [9, 18]. A common practice leverages weighted schemes to construct a linear combination of the input networks, by computing for the disease under study an informativeness coefficient for each network. For instance, in [13] the informativeness of a network has been computed as the percentage of decay in the area under the ROC curve or under the precision-recall curve of a given classifier when removing that network from the integration process. We show in this study that such a coefficient should take into account the rarity of known disease-genes characterizing most entries 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). Indeed, when a disease-gene (positive gene) is rare for a given disease, it carries most information about the latter, and in principle an input source should be considered informative when it embeds information (in the form of gene connections) allowing a given classifier to correctly rank positive genes. Such an integration process is usually called “imbalance-aware”, and it already led to successful results in similar contexts, such as the protein function prediction [8]. Unfortunately, the central issue represented by the rarity of disease-genes has been neglected by most existing approaches for data source integration for gene prioritization.

We argue that network integration must be imbalance-aware even for the GP problem, to improve the accuracy of gene rankings. To this purpose, we leveraged a method recently proposed for imbalance-aware integration in the context of protein function prediction, *UNIPred* (Unbalance-aware Network Integration and Prediction, [8]), and extended it in order to emphasize the important role disease-genes play in the integration process. Informally, UNIPred operates a projection of the network onto the plane, where the projected points/genes constitute the items of a new optimization problem, whose solution provides the informativeness coefficient for the input network (see [8] for theoretical details). This method has been extended by introducing a novel optimization criterion, in which the relevance to be attributed to disease-genes is associated with a free parameter, so as to easily verifying our claim. Through the network usefulness computed by UNIPred, the consensus network is built and given as input to *WGP*, a recent network-based algorithm proposed to prioritize disease-genes [6]. The overall methodology has been then validated on a benchmark data set composed of nine human networks and 708 MeSH disease terms [18].

¹ <http://www.nlm.nih.gov/mesh/>.

2 Materials

Our setup follows a benchmark proposed in [18] 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 [18] for details about each network).

Functional interaction network – *finet*. A network covering 8441 selected proteins and containing protein–protein functional binary interactions predicted through a Naive Bayes classifier trained on a ‘gold’ pairwise relationships set extracted from curated pathways [19].

Human net – *hnnnet*. 21 large-scale genomics and proteomics data sets from human and from orthologs in yeast, fly and worm are integrated by including distinct lines of evidence, spanning human mRNA co-expression, protein-protein interactions, protein complex, and comparative genomics data sets [10].

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 gene–chemical interactions available at the Comparative Toxicogenomics Database (CTD) [4].

BioGRID database network – *dbnet*. BioGRID protein–protein interaction network for 8449 proteins based upon direct physical and genetic interactions constructed in [18].

BioGRID projected network – *bgnnet*. 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, [1]) terms in the three branches annotating the considered genes: biological process (bp), molecular function (mf) and cellular component (cc). The connection between two genes is given by the maximum Resnik semantic similarity between all the terms (in that branch) the two genes are annotated with.

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

A network integration problem assumes m network sources about gene pairwise similarities are given, every source represented through a weighted undirected graph $G^{(k)} = \langle V, \mathbf{W}^{(k)} \rangle$, where V is the set of genes/instances (or a subset of it), $k \in \{1, 2, \dots, m\}$ is the network index and $\mathbf{W}^{(k)}$ is the connection matrix: the

entry $W_{i,j}^{(k)} \in [0, 1]$ indicates a degree of functional similarity between genes i and j . If a data source covers just a subset of genes in V , we extended it to V by adding zeros in the corresponding entries of its connection matrix. We assume thereby in the following that all networks cover the set V . Given a disease of interest d , every gene $i \in V$ is associated with 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 $G_d = \langle V, \mathbf{W} \rangle$ integrating all available networks, to be used to predict candidate disease-genes for d . This is performed by associating every network $G^{(k)}$ with a coefficient $r_d^{(k)}$ related to its informativeness for disease d , and then by linearly combining input networks through the obtained coefficients (see Sect. 3.2). To compute $r_d^{(k)}$ we adopt an extension of the UNIPred algorithm, briefly described in the following.

3.1 UNIPred

The UNIPred algorithm computes for every networked-source $G^{(k)}$ 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 $P_i^{(k)}$, embedding the local imbalance in the corresponding node position. The coordinates $P_i^{(k)} \equiv (P_{i,1}^{(k)}; P_{i,2}^{(k)})$ are computed as follows:

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

In other words, $P_{i,1}^{(k)}$ is the weighted sum of 1-labeled neighbors, $P_{i,2}^{(k)}$ 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 (Fig. 1).

The algorithm then learns the straight line which best separates positive and negative points, in the sense we describe below. Since every point $i \in V$ already has a label y_i , each line separating positive and negative points is associated with the number $TP_d^{(k)}$ of positive points correctly classified (true positives), the number $FN_d^{(k)}$ of positive points wrongly classified (false negatives), and the number $FP_d^{(k)}$ of negative points wrongly classified (false positives). The optimal line is the one maximizing the F -measure: $F_d^{(k)} = \frac{2TP_d^{(k)}}{2TP_d^{(k)} + FP_d^{(k)} + FN_d^{(k)}}$. The value $\bar{F}_d^{(k)}$ corresponding to the optimal line is then considered as relevance $r_d^{(k)}$ for the input network $G^{(k)}$. The method is imbalance-aware since the F -measure by definition penalizes more heavily the misclassification of positive instances, with

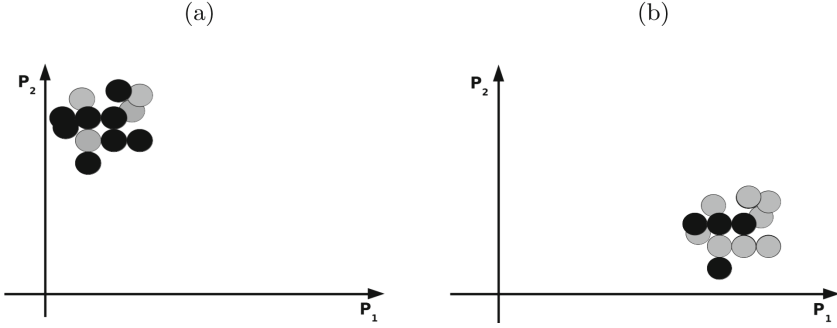


Fig. 1. Examples of distributions of points $P_i^{(k)}$ for a given network $G^{(k)}$ in which labels are unbalanced towards (a) negative points (black) and (b) positive points (light grey). In the case (a), the coordinate P_2 tends to be much larger due to the predominance of negative neighbours; as opposite, P_1 is larger in case (b), since the large majority of neighbours in average is positive.

respect to the penalty for misclassifying negatives. Moreover, maximizing $F_d^{(k)}$ moves the know labeling $\mathbf{y} = (y_1, \dots, y_{|V|})$ towards a minimum of the energy of underlying Hopfield network—allowing the model to better fit the input data (see [8]).

In order to emphasize the need of attributing higher importance to positive genes, here we extend UNIPred by adopting the variant F_β of F , defined as it follows:

$$F_{\beta,d}^{(k)} := \frac{(1 + \beta^2)TP_d^{(k)}}{(1 + \beta^2)TP_d^{(k)} + FP_d^{(k)} + \beta^2FN_d^{(k)}}. \quad (2)$$

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 assign a higher penalty to the misclassification of positives. The larger β , the more relevant are positives in determining the network coefficient $r_d^{(k)}$. Since β is dependent on the input data, we learned it through internal cross validation in our experimentations; in Sect. 4 we also supply the results of tuning β , to investigate its impact on the algorithm performance.

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 $\mathbf{r}_d = (r_d^{(1)}, r_d^{(2)}, \dots, r_d^{(m)})$. The consensus network is then constructed as a weighted sum (WS) of the corresponding adjacency matrices:

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

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

3.3 Inferring the Gene Prioritization List

Once the consensus network $G_d = \langle V, \mathbf{W} \rangle$ for disease d 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 , disjoint subsets of V respectively containing the labeled and unlabeled genes, and the objective is to infer a ranking of genes in U with respect to d . Only for genes $i \in L$ the label $y_i \in \{0, 1\}$ is thereby known, and the aim is learning a function $\phi : U \rightarrow \mathbb{R}$ 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 when the imbalance between positive and negative genes is large. Accordingly, the adopted methodology has to consider this feature of the problem to prevent a large decay of the ranking quality [5]. To learn the ranking function ϕ we employed a regression model proposed in [6], termed *WGP (Weighted Gene Prioritization)*, able in handling the label imbalance during the prioritization process. Briefly, starting from the integrated network, WGP learns a weighted binomial regression model with log-log link function, a skewed function suitable for unbalanced data, to separate positive and negative nodes, and consequently infer the prediction for genes U using the learned regression model.

4 Results

Following the benchmark setting [18], the generalization performance of our method has been assessed through a classical 5-fold cross-validation procedure, and the results have been evaluated by using the Area Under the Receiver Operating Characteristic Curve (*AUC*) and the Precision at different Recall levels (*PxR*). In addition, we have 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 obtained results on benchmark data show a noticeable and statistically significant improvement of validation of WGP-UNIPred algorithm with respect to the compared methods (Wilcoxon signed rank test, *p-value* < 0.01), including random walks [11], random walks with restarts, guilt-by-association methods [12] and kernelized average score functions (S_{AV} [17]). In particular S_{AV} , the top benchmark method, is based on an extension of the gene–gene similarity to non neighboring nodes by adopting a suitable kernel matrix. The score for each gene i with regard to a given disease d is defined according to a suitable distance $d(i, V_d)$ between i and the subset V_d of genes positive for d . In S_{AV} , $d(i, V_d)$ is defined as the average distance between the images in the corresponding Hilbert space of i and the elements in V_d (see [17] for details).

Figure 2 shows the overall performance, remarking both the gain of UNIPred with respect to US integration scheme and the influence of the β parameter on the performance. We only report the results of S_{AV} with weighted and unweighted sum integration, since random walk and the other compared methods achieved worse results than S_{AV} . In [18], the average AUC results across diseases have been used to weight networks according to the WS integration for S_{AV} . The β parameter has been tuned in the set of values $\{1, 2, 3, 4, 5, 10, 15, 20, 25\}$, in this first experiment, to show how it influences the model performance. 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 schemes, there is an almost negligible difference between *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)

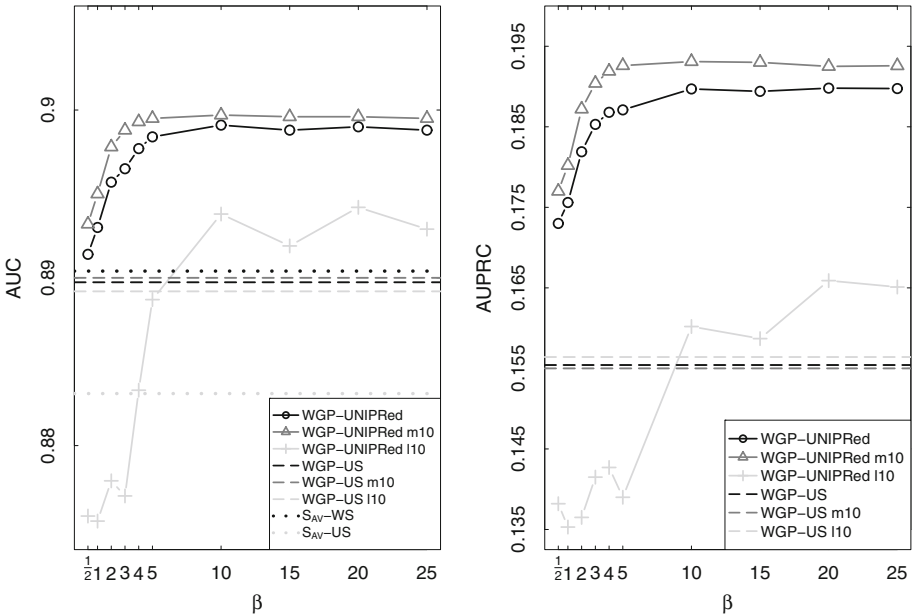


Fig. 2. Performance of WGP-UNIPRED on benchmark data. ‘l10’ and ‘m10’ refer to the subsets of MeSH disease terms with 5–10 and 11–200 associated genes, respectively, whereas circles correspond to results averaged across all diseases. WGP-US is the average performance across all diseases of WGP on unweighted sum data, whereas WGP-US l10 (resp. WGP-US m10) denotes the WGP performance on US data averaged across the category ‘l10’ (resp. ‘m10’). S_{AV} results are averaged across all diseases.

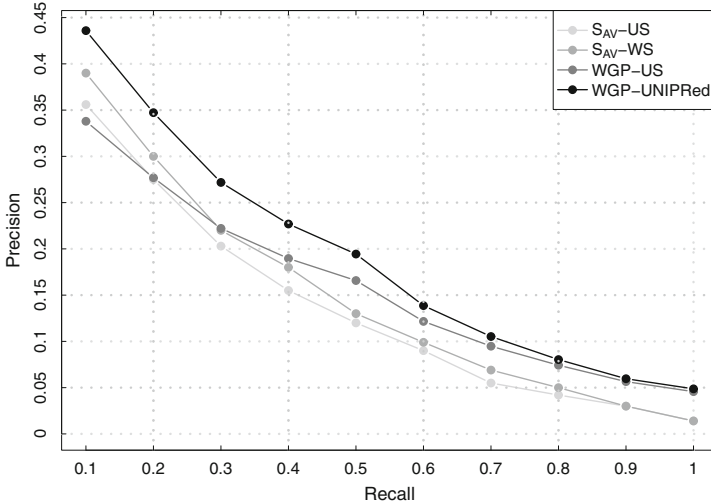


Fig. 3. PxR results achieved by the top benchmark method S_{AV} and WGP-UNIPred on both unweighted and weighted schemes.

to weight networks is more effective in this context. Apparently, the larger improvement for UNIPred compared to US scheme 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 leads to better results [14]. We also compared the methods in terms of PxR (Fig. 3): in this experiment we learned β through internal cross validation. WGP-UNIPred favourably compares even in terms of PxR, outperforming S_{AV} in all experiments and WGP-US in all but 0.1 recall settings, where results are almost indistinguishable. Confirming the behaviour in terms of AUC, the UNIPred weighted sum integration led to larger improvements (mainly for lower values of recall) than the imbalance-unaware weighted integration of S_{AV} , with regard to the US corresponding results.

5 Conclusion

Experimental results supported our claim that the integration of omics data (genomics, transcriptomics, proteomics and so on) needs imbalance-aware procedures for improving the accuracy of gene prioritization lists. A state-of-the-art integration algorithm, UNIPred [8], has been used to boost the performance of a gene prioritization method, WGP [6]. 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 funded grant title *Machine learning algorithms to handle label imbalance in biomedical taxonomies*, code PSR2017_DIP_010_MFRAS, Università degli Studi di Milano.

References

1. Ashburner, M., et al.: Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**(1), 25–29 (2000)
2. Barabasi, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**(1), 56–68 (2011). <https://doi.org/10.1038/nrg2918>
3. Che, J., Shin, M.: A meta-analysis strategy for gene prioritization using gene expression, SNP genotype, and eQTL data. *BioMed Res. Int.* **2015**, 1–8 (2015). <https://doi.org/10.1155/2015/576349>
4. Davis, A.P., et al.: Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.* **37**(Database issue), D786–D792 (2009). <https://doi.org/10.1093/nar/gkn580>
5. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973–978 (2001)
6. Frasca, M., Bassis, S.: Gene-disease prioritization through cost-sensitive graph-based methodologies. In: Ortuño, F., Rojas, I. (eds.) IWBBIO 2016. LNCS, vol. 9656, pp. 739–751. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31744-1_64
7. Frasca, M.: Gene2DisCo: gene to disease using disease commonalities. *Artif. Intell. Med.* **82**, 34–46 (2017). <https://doi.org/10.1016/j.artmed.2017.08.001>
8. Frasca, M., Bertoni, A., Valentini, G.: UNIPred: Unbalance-aware Network Integration and Prediction of protein functions. *J. Comput. Biol.* **22**(12), 1057–1074 (2015). <https://doi.org/10.1089/cmb.2014.0110>
9. Frasca, M., Malchiodi, D.: Exploiting negative sample selection for prioritizing candidate disease genes. *Genomics Comput. Biol.* **3**(3), e47 (2017). <https://doi.org/10.18547/gcb.2017.vol3.iss3.e47>
10. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., Marcotte, E.M.: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**(7), 1109–1121 (2011). <https://doi.org/10.1101/gr.118992.110>
11. Lovász, L.: Random walks on graphs: a survey. In: Miklós, D., Sós, V.T., Szónyi, T. (eds.) *Combinatorics, Paul Erdős is Eighty*, vol. 2, pp. 353–398. János Bolyai Mathematical Society, Budapest (1996)
12. Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., Eisenberg, D.: A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999)
13. Montojo, J., Zuberi, K., Shao, Q., Bader, G.D., Morris, Q.: Network assessor: an automated method for quantitative assessment of a network’s potential for gene function prediction. *Front. Genet.* **5**, 123 (2014). <https://doi.org/10.3389/fgene.2014.00123>
14. Mostafavi, S., Morris, Q.: Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* **26**(14), 1759–1765 (2010)
15. Piro, R.M., Di Cunto, F.: Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.* **279**(5), 678–696 (2012). <https://doi.org/10.1111/j.1742-4658.2012.08471.x>

16. Tiffin, N., Andrade-Navarro, M.A., Perez-Iratxeta, C.: Linking genes to diseases: it's all in the data. *Genome Med.* **1**(8), 77 (2009). <https://doi.org/10.1186/gm77>
17. Valentini, G., et al.: RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics* **32**, 2872–2874 (2016). <https://doi.org/10.1093/bioinformatics/btw235>
18. Valentini, G., et al.: An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif. Intell. Med.* **61**(2), 63–78 (2014). <https://doi.org/10.1016/j.artmed.2014.03.003>
19. Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**(5), 1–23 (2010). <https://doi.org/10.1186/gb-2010-11-5-r53>