Connectivity significance for disease gene prioritization in an expanding universe

Manuela Petti, Daniele Bizzarri, Antonella Verrienti, Rosa Falcone and Lorenzo Farina

Abstract— A fundamental topic in network medicine is disease genes prioritization. The underlying hypothesis is that disease genes are organized as modules confined within the interactome. Here, we propose a novel algorithm called DiaBLE (**DIA**MOnD's **B**ackground **L**ocal **E**xpansion) which is a modified version of DIAMOnD, a successful algorithm based on the concept of connectivity significance. Instead of taking the whole interactome as the background model, DiaBLE considers as gene universe the smallest local expansion of the current seeds set at each iteration step. We show that DiaBLE significantly increases the overall DIAMOnD ranking quality of genes prioritization both in terms of cross-validation and biological consistency. Here, we focus on the two algorithms only since a comparative analysis among gene prioritization methods is beyond the scope of this study. Finally, we briefly discuss the improvement of biological insight provided by DiaBLE for two cancers (head and neck squamous cell carcinoma and kidney renal clear cell carcinoma).

Index Terms—Network problems, distribution functions, Biology and Genetics, Molecular biology, Life and medical sciences

1 Introduction

TODAY, big data, genomics, and quantitative *in silico* integration methodologies, have the potential to push forward the frontiers of medicine in an unprecedented way [1], [2]. A large body of evidence that is now emerging from new genomic technologies, points out directly to the cause of disease as perturbations within the interactome, *i.e.* mutations potentially impacting the comprehensive network map of molecular components and their interactions [1]. As a matter of fact, a fast growing experimental evidence reveals the association between groups of interacting proteins and disease within the human interactome, representing the cellular network of all physical molecular interactions [3]. Precisely, the human interactome is composed of direct physical, regulatory (transcription factors binding), binary, metabolic enzyme-coupled, protein complexes and kinase/substrate interactions. Such network is largely incomplete as well as the connections between genes and disease.

Disease proteins are the product of genes whose mutations have a causal effect of the respective phenotype. A key property of the underlying molecular network of interactions is that disease proteins are not found to be uniformly scattered across the interactome, but they tend to interact with one another confined in one or several subgraphs called "disease modules" [4]. In fact, disease proteins are prone to participate in common biological activities such as, for example, genome maintenance, cell differentiation or growth signaling, which are the most relevant in carcinogenesis [5]. Consequently, the module property also reflects the biological feature that disease proteins are often localized on specific biological compartments (pathway, cellular space, or tissue). These considerations directly point towards the possibility that, whenever a disease module sub-network is found, other disease-related parts are likely to be identified in their topological neighborhood [3]. However, notwithstanding a strong community commitment to find new protein interactions and relevant mutations for disease characterization, the list is still largely incomplete. Moreover, identification of specific disease genes is often impaired by gene pleiotropy, by the multigenic feature of many diseases, by the influence of a plethora of environmental agents, and by genome variability [6]. The problem that we tried to address in this paper is the prioritization of candidate disease genes, *i.e.* a computational approach to ranking for their potential as genes harboring disease-driving mutations. The knowledge gained through these types of analysis could help the understanding of the pathogenesis processes and the identification of cellular molecular profiles providing a useful tool in improving diagnosis, prognosis and therapy. Indeed, this problem has motivated the development of a number of algorithms [7]. The key question is to find a way to fully characterize such genes (with respect to non-disease genes) and find an algorithm able to capture such characteristics. From a network perspective, one hypothesizes that disease genes are embedded within modules in ways that are amenable to some topological feature description. The recent [4] evidence-based biological observation that disease genes are not randomly positioned in the interactome has opened new possibilities for developing algorithms for disease gene predictions.

- *M.P. is with the Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy. E-mail: manuela.petti@uniroma1.it.*
- *D.B. is with the Department of Translational and Precision Medicine, Sapienza University of Rome, Italy. E-mail: deniele.bizzarri@gmail.com.*
- *A.V. is with the Department of Translational and Precision Medicine, Sapienza University of Rome, Rome, Italy. E-mail: antonella.verrienti@uniroma1.it.*
- *R.F. is with the Department of Translational and Precision Medicine, Sapienza University of Rome, Rome, Italy. E-mail: rosa.falcone@uniroma1.it.*
- *L.F. is with the Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy. E-mail: lorenzo.farina@uniroma1.it. (Corresponding Author)*
- *M.P. and D.B. contributed equally to this work*

Two groups of methodologies have emerged in the last decade as the most promising ones: network propagation [8] and modules-based [3], [7] algorithms. Network propagation (or diffusion-based) algorithms rely on the assumption that the information contained in the initial (known) set of disease genes, flows through the network through nearby proteins: among all, one of the most used diffusion-based algorithms is the "random-walk with restart" originally proposed by Köhler *et al.* in [9]. By contrast, the so-called "module-based algorithms" rely on the hypothesis that all cellular components that belong to the same topological, functional or disease module have a high likelihood of being involved in the same disease. In [10] the authors compare different approaches based on different strategies (*e.g.* network neighbors [11], clustering [12], random walk [9], propagation [13]) for linking genes with diseases and conclude that, although random-walk based algorithms individually outperform clustering and neighborhood approaches, all the methods provide relevant biological insight by exploiting different network properties. A review of methods has been presented by Moreau and Tranchevent [14].

From the above introductory discussion, prioritizing candidate disease genes using the interactome, *i.e.* the network of physical protein interactions, and mutational data (known disease gene or seeds), is still an open problem. In fact, diseases are highly heterogeneous and the topological patterns of its mutated genes on the relevant interactome (the disease module) is also highly heterogeneous. Therefore, it is unreasonable to believe that a single algorithm could be able to "catch" all the biological information contained in the data and be the "best one" or "state of art" for any possible disease.

A reliable prioritization (or ranking) of new predicted disease genes is very important from a biological viewpoint, since it provides valuable information of a putative specific activity of a gene in the development of a disease. In fact, the smaller its rank position, the more likely it is a "true" disease gene, thus providing an ordered list to the experimenter/clinician who can decide which are the most promising candidates for experimental testing.

Here, we deal with the problem of prioritizing disease genes using the concept of connectivity significance recently introduced by Ghiassan *et al.* [7] in their successful module-based algorithm called DIAMOnD [7]. The algorithm assumes the availability of interactome data (network of protein-protein physical interaction) and seed genes (known disease-associated genes). The methodology is based on the key observation that putative disease genes exhibit distinct and predictive connectivity patterns on the interactome. Such network-based signatures can be captured and exploited if one evaluates the significance of their connections instead of their density, *i.e.* by using the concept of *connectivity significance* [7] as defined in the next section. Another important feature of the DIAMOnD algorithm is that it reduces the spurious detection of high degree genes [7].

Here, we fully support the underlying rationale of Ghiassan *et al.* methodology [7]. This approach - in our opinion - has been the key to success for effective disease genes detection. For this reason, we will further pursue this pivotal idea and suggest a modified version of the DIAMOnD algorithm called DiaBLE (**Dia**mond **B**ackground **L**ocal **E**xpansion) that introduces a new connectivity significance score by considering an adaptive gene universe in the associated hypergeometric test. A comparative analysis among gene prioritization algorithms based on different strategies is beyond the scope of this study. As a final note, we recall that the interactome is made of proteins, but we often talk about disease genes. To avoid terminological confusion, in the following, we will refer only to disease genes, unless the intended meaning is not clear from the context.

2 Method

We begin this section by first illustrating the DIAMOnD algorithm procedure [7] so to clarify the reasoning underlying the modification introduced by the DiaBLE algorithm, as fully described in the subsequent paragraph.

2.1 The DIAMOnD algorithm: a stationary universe

An important observation leading to the development of the DIAMOnD algorithm by Ghiassan *et al.* [7] is that topological communities are not able to capture disease modules, which are not always organized as dense clusters. However, disease genes of a specific module exhibit a peculiar topological pattern characterized in terms of connectivity significance rather than density. Based on this key idea, the DIAMOnD algorithm defines a *connectivity P-value* of a gene as follows:

$$p - value = \sum_{k_i=k_s}^{k} p(k, k_i) \qquad (1)$$

with

$$p(k, k_s) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}} \qquad (2)$$

which is the probability that a gene with a total of k links has k_s connections to seeds (set composed of s_0 elements) given a network of N genes. The underlying null hypothesis (background model) is that seed genes are randomly scattered throughout the whole interactome. Consequently, the gene universe is stationary (fixed) and composed of N nodes.

Having in mind definition (1), we can state the DIAMOnD algorithm [7]. The steps required to infer new putative disease genes are the following:

i) Determine the connectivity significance (1) for all M genes linked to any of the s_i (starting from s_0) seed genes.

ii) The genes are ranked according to their respective *P*-values (ascending order).

iii) The gene with the highest rank (lowest *P*-value) is added to the set of seed nodes, increasing their number from $s_i \rightarrow s_{i+1} = s_i + 1$.

iv) Steps *(i)-(iii)* are repeated with the expanded set of seed genes, pulling in one gene at a time into the growing disease module (seeds set).

The genes defined at step i) (*i.e.* seed neighbors) are called "candidate genes". The procedure can be continued until the DIAMOnD module spans across the entire network. However, the expected dimension of the disease module is usually 200-400 proteins [7]. Candidate genes included in the growing disease module identified by the algorithm starting from the initial seed set, will be called "DIAMOnD genes". Analogously, we will define "DiaBLE genes".

Finally, we note that at each iteration step defined by i), a *P*-value must be computed for each of the *M* candidates, *i.e.* genes of the interactome having at least a link to genes of the current seed set s_i, at iteration *i*. Each single *P*-value computation requires the selection of a candidate gene having $k_s \geq 1$ links to the current seed set of size s_i, *k* first neighbors and a gene universe composed of *N* elements. The one having minimal *P*-value is considered the "best" candidate, added to the seed set, and the procedure con move forward to the subsequent iteration step. It is worth noting that DIAMOnD assumes a fixed gene universe which *coincides with the whole interactome*.

2.2 The DiaBLE algorithm: an expanding universe

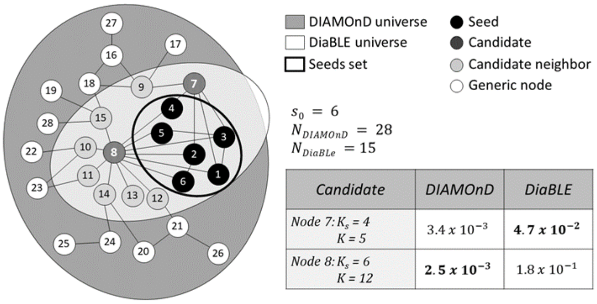
Generally speaking, to perform a hypergeometric test [15], [16], one needs to define a gene universe (for example, all the genes of the interactome, as in [7]) and a selection from that universe (seed genes). The subsequent step is the identification of the subset of the universe that is considered "interesting" (new putative disease genes, identified by having at least one link to the seed set). As discussed in full detail by Falcon and Gentleman [15], [16], the selection of the universe is very important, since it has a large impact on the observed *P*-values. The recommendation of the authors is "*to include in the universe only those genes that could have been selected as interesting*" [16]. It is clear that the choice of the set of potentially interesting genes is highly subjective (an *a priori*) and it makes no sense to consider one as better than another. The DIAMOnD algorithm considers *all* the genes in the interactome as potentially interesting, which a reasonable choice, but not necessarily the only one.

The basic idea underlying the DiaBLE algorithm is to explore the impact of a different gene universe selection, at each iteration step *i*. Here, we propose to choose, starting from the current seed set s_i, its smaller expansion, *i.e.* a sort of local universe, as opposed to the global universe of all the genes in the interactome. To obtain such local universe set we considered as "*genes that could have been selected as interesting*" [16] only the *M* candidate genes and their first neighbors. In other words, the gene universe considered by DiaBLE, is the union of:

i) the current seed set
ii) the *M* candidate genes (*i.e.* those having at least a link to the current seed set)
iii) the first neighbors of the candidate genes

In this way, all the values in formula (1) can be computed and are actually the same as DIAMOnD, except for the universe size *N* which is now an increasing number depending on the iteration step of the algorithm. The DiaBLE universe is the "smallest" universe expansion of the seeds set at iteration *i* from which it is possible to obtain the values (*k*, k_s, s_i, N) needed to compute *P*-values defining the connectivity significance.

The effect of this local universe on the choice of the "best" candidate gene can be critical, since not only the *P*-values but also the ranking of the *M* candidates can be deeply modified, as illustrated by a toy example in Figure 1.



Candidate	DIAMOnD	DiaBLE
Node 7: $K_s = 4$ $K = 5$	$3.4 \, x \, 10^{-3}$	$\mathbf{4.7 \, x \, 10^{-2}}$
Node 8: $K_s = 6$ $K = 12$	$\mathbf{2.5 \, x \, 10^{-3}}$	$1.8 \, x \, 10^{-1}$

Fig. 1. A toy example of a generic *i*-th step of the algorithm, illustrating the differences between DIAMOnD and DiaBLE. As shown by the table in the figure, the size of the universe considerably impacts the *P*-values and, most importantly, their ranking. In fact, according to DIAMOnD, the best candidate is node 8, whereas DiaBLE would choose node 7. Note that node 8 has a much higher degree than node 7 and that 90% of the links from node 7 point to seed nodes, whereas only 50% of the links from node 8 do the same. This toy example also shows the ability of DiaBLE to reject spurious high degree nodes.

Now, we can describe the DiaBLE algorithm to infer new disease genes. The required steps are the following:

i) Determine the universe size N_i by considering the set composed of: seed genes, genes that have at least one link to the seed set (candidate genes) and their first neighbors.
ii) Determine the connectivity significance (1) for all *M* genes linked to any of the s_i (starting from s_0) seed genes (called "candidate genes") using N_i in formula (2) instead of *N*.
iii) The genes are ranked according to their respective *P*-values (ascending order).
iv) The gene with the highest rank (lowest *P*-value) is added to the set of seed nodes, increasing their number from $s_i \rightarrow s_{i+1} = s_i + 1$.
v) Steps *(i)-(iv)* are repeated with the expanded set of seed genes, pulling in one protein at a time into the growing seeds set (or module).

It is plain that the only difference between DIAMOnD and DiaBLE is the size of the universe set used to compute connectivity significance at each step *i*. Nevertheless, in what follows we show that this simple modification of the universe has, surprisingly, a significant impact on the disease module composition. Most importantly, its impact on performances using computational and biological validation sets is the presence of a significant increase.

3 Results and Discussion

Here we show – with no additional computational efforts – that DiaBLE improves the overall DIAMOnD performances in a statistically significant way. We used the same computational and biological benchmarks as in [7], to make the comparison fair. Moreover, we will show that DiaBLE is also effective – just like DIAMOnD – in reducing the impact of spurious detection of high degree genes. To evaluate and compare performances of DiaBLE and DIAMOnD algorithms, we used the same seed genes associated to 70 diseases and the same human interactome used by Ghiassan *et al.* in reference [7]. The resulting network is composed of 13,460 proteins interconnected by 141,296 physical interactions: this set of physical interactions is the result of the integration of several databases as explained in detail in reference [7]. To compare performances of DiaBLE and DIAMOnD, we also defined validation criteria. Following Ghiassan *et al.* [7], we firstly considered a cross-validation approach and, secondly, a biological criterion based on gene annotations (obtained from GO and KEGG databases), as defined in [7]. The underlying rationale of this criterion is that a "true" new disease gene should share at least one biological annotation with the original seed set. Quality of ranking is therefore evaluated according to the presence of common annotations along subsequent iterations. Clearly, at each iteration, a potentially new disease gene is predicted by an algorithm and the higher the position in the rank, the better it performs.

3.1 Resilience to spurious detection of high degree nodes

It is worth showing an interesting preliminary difference between the two methodologies. We selected top 50 DiaBLE genes and top 50 DIAMOnD genes lists for all the 70 diseases and compared them in terms of the degree (k) and the ratio k_s/k (a measure of the communication level from a gene/node towards the seeds). For each disease and algorithm, we computed the average value of k and k_s/k across the selected genes. Figure 2 shows the results: DiaBLE genes tend to have, on the one hand, higher values of the ratio k_s/k (Wilcoxon signed rank test P-value $p = 1.52 \cdot 10^{-4}$), *i.e.* a higher fraction of links connected to the seed set, and, on the other hand, less nodes with a large number k of links (Wilcoxon signed rank test P-value $p = 5.47 \cdot 10^{-7}$). Moreover, if we consider the first 500 genes generated by both algorithms, the statistically significant differences related to the ratio k_s/k does not hold anymore, whereas the global higher degree k of DIAMOnD genes still holds (Wilcoxon signed rank test P-value $p = 2.47 \cdot 10^{-10}$). Taken together, these results, show that DiaBLE, on average, privileges higher values of the ratio k_s/k with smaller values of k, thus showing to be more resilient to spurious detection of high degree nodes. To evaluate whether DiaBLE privileges also the selection of seed neighbors, we calculated for each disease the number of DIAMOnD and DiaBLE genes directly connected to the seed set. We did not observed a significant difference between the two algorithms in terms of the amount of selected seeds' neighbors (figure S1); however, the higher Δ_{NN} value is related to the 70th disease (vasculitis), case in which 141 DIAMOnD genes are directly connected to the seed set, while DiaBLE selects only 57 seed neighbors out of 500 predicted genes.

3.2 Cross validation (out-of-sample test)

As stated in the previous paragraph, following [7], the first validation criterion we considered was the cross-validation approach, or out-of-sample testing. The goal is to test the ability of the algorithms to recover the original seed set, starting from a given percentage (usually, 90%, 80% and 70%) obtained by random seed gene selection. The procedure must be repeated a certain number of times so that a statistical evaluation of performances can be reliably computed. Accordingly, we randomly selected (removed) genes at various percentages (10%, 20% and 30%) and, for each disease, we computed the average recall (or true positive rate) obtained from 50 replicates. Then, we averaged (median) the results across diseases and plotted the resulting recall curve for both methods. To compute statistical differences between DiaBLE and DIAMOnD curves, we accumulated recall vectors of all diseases in two separate vectors (one for each algorithm) and performed a two-tailed Wilcoxon signed rank test on the difference vector. The results are presented in Figures 3A and 3B, where it is clearly shown that DiaBLE performs better than DIAMOnD for all percentage removals (Wilcoxon signed rank test P-value < 10^{-40}) over all the 70 diseases. It is very worth noting that, as shown by Figure 3B, the increase in performance begin to be visible starting from about iteration 20/30, thus confirming the more efficient prioritization generated by DiaBLE in presenting "true" seed genes at the very top of the ranked list.

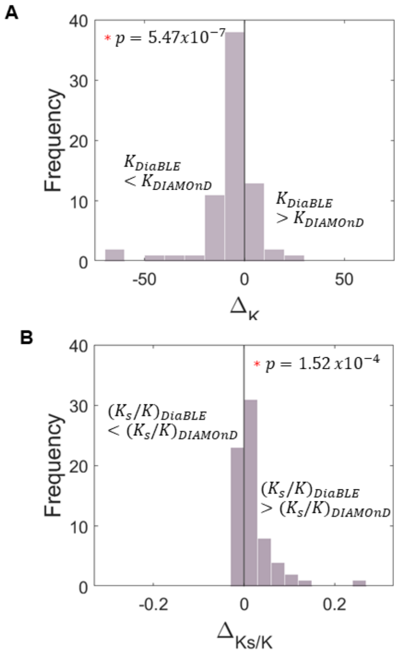


Fig. 2. **A)** Histogram of the differences $\Delta_{K_i} = k_{DiaBLE_i} - k_{DIAMOnD_i}$ with $i = 1, \dots 70$ and **B)** histogram of the differences $\Delta_{(K_s/K)_i} = (k_s/k)_{DiaBLE_i} - (k_s/k)_{DIAMOnD_i}$ with $i = 1, \dots 70$ with the associated *P*-value returned from the Wilcoxon sign rank test. The diagrams make clear that the distribution of the rations k_s/k for DiaBLE genes show more diseases with higher fraction of links to the seeds set and a reduction of spurious nodes, *i.e.* those having a high degree k

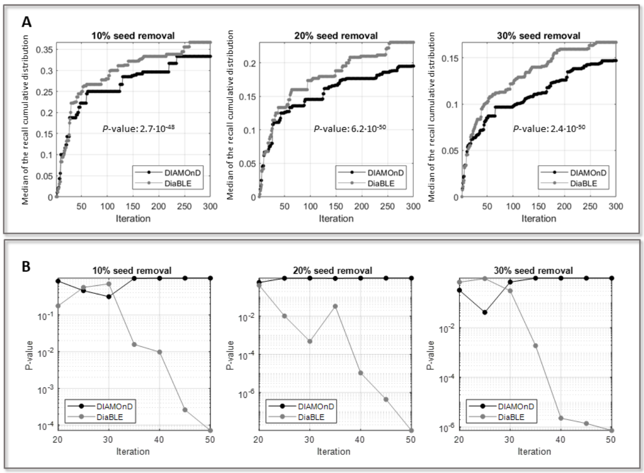


Fig. 3. **(A)** Median of the cumulative recall distribution across the 70 diseases, for DiaBLE and DIAMOnD genes, in the cases of 10% (left panel), 20% (central panel) and 30% (right panel) nodes removal. Wilcoxon signed rank test reveals in all cases a highly significant difference of the median values (*P*-value <10^{-40}) in favor of DiaBLE. (**B**) *P*-values corresponding to the initial 50 iterations (Wilcoxon signed rank test). It is worth noting that increased overall performances are visible starting from iteration 20/30, thus confirming the positive impact of the choice of an expanding gene universe in the top "new disease" genes.

To further study the improvement of DiaBLE in prioritizing disease genes, we performed a statistical comparison (Wilcoxon signed rank test, *P-value* < 10^{-3}) between the two algorithms by considering recall vectors of each disease. In particular, we performed the comparison by taking into account different iteration values: for each disease, the first test was performed by considering the first 25 iterations with the aim to reveal differences related to the top candidates, then we considered different iteration values by increasing their number with a step of 25. Figure 4 shows the results obtained. For each of iteration value ($N_{it} \in [25; 300]$), the bar diagrams (panels *a*, *b* and *c* for the cases of 10%, 20% and 30% nodes removal, respectively) show the number of diseases for which the performance are better using DIAMOnD (black bars) and DiaBLE (grey bars). It is worth noting that, since the initial iteration values (*i.e.* considering the first 20/30 iterations), DiaBLE outperforms DIAMOnD in a larger number of cases. A detailed representation of the case of 30% node removal is shown in figure 4.d, where each row represents one of the 70 diseases (see Supplementary Tables S1 for matching numbers to diseases). The figure shows the performance obtained considering each specific disease and informs that in very few cases, there is a reversal of performance in favor of one algorithm with respect to the other.

3.3 Biological validation results (GO terms and pathway enrichment)

Disease gene prediction is obtained by an iterative procedure, as described in the Methods section, so that at each iteration a different "new" gene is added to the seed set depending on the algorithm (a DIAMOnD gene or a DiaBLE gene). Following Ghiassan *et al.* [7], a "new" disease gene (DiaBLE or DIAMOnD) has been considered a *true*

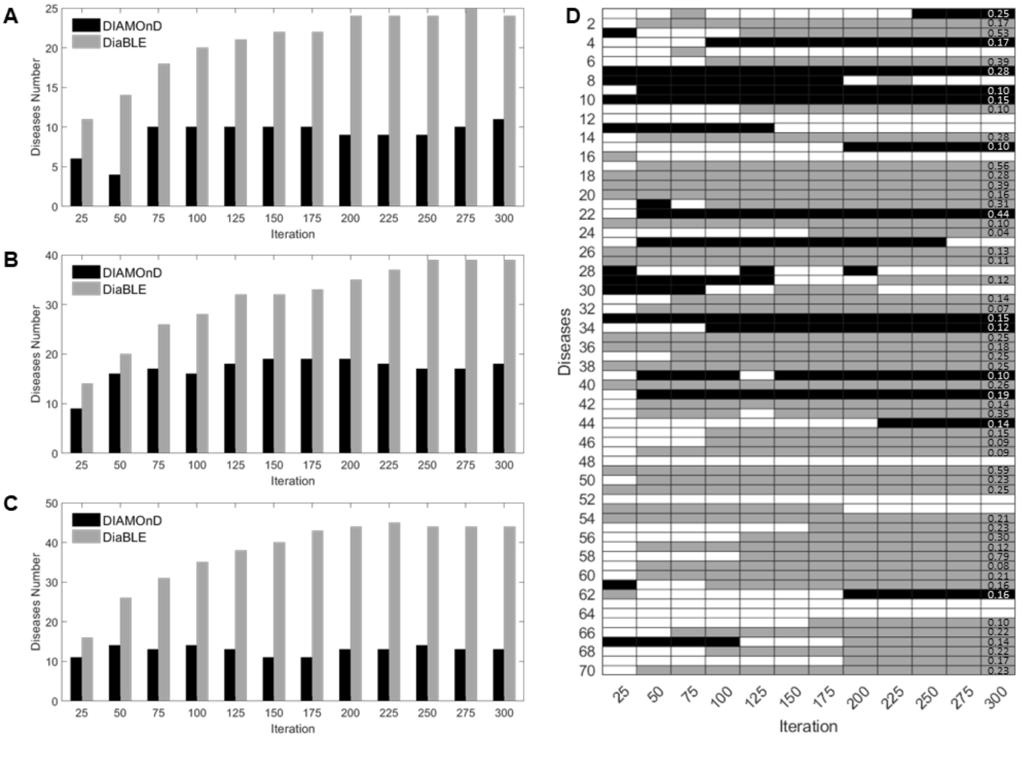


Fig. 4. Bar diagrams (panels **A**, **B** and **C** for the cases of 10%, 20% and 30% nodes removal, respectively) showing the number of diseases for which the performances are statistically better (Wilcoxon signed rank test, *P-value* < 0.001) using DIAMOnD (black bars) and DiaBLE (grey bars). Panel **D** shows a detailed representation of the case of 30% node removal with the recall values at 300 iterations (reported only when one of the two algorithms outperforms the other.

hit (TH) if it is characterized by sharing at least, in the corresponding database (GO or KEGG), one of the most significant annotations of the seeds set. The rationale behind this validation procedure is that it is assumed that the original seeds set is an un-biased sample of the complete seeds set. Hence, a new disease gene is more likely to be "true" if it has a physical interaction with the seeds (*i.e.* a link on the interactome) and, if it participates in the same biological process of – at least – a gene belonging to the current seeds set. Hence, the first step is to find significant enrichments of a given seeds set using biological annotations. We therefore considered the annotations provided by GO terms (Gene Ontology database, biological process, downloaded April 2018) and pathways (KEGG gene set from the Molecular Signatures Database, downloaded April 2018). The available GO terms (biological process), were prefiltered as follows [7]:

i) annotations labeled with evidence code "IPI" (Inferred from Physical Interaction) were excluded to avoid circularity;
ii) annotations not associated with the gene products (evidence code "NOT") were excluded.

Moreover, to obtain smaller gene lists, annotations were not propagated upwards on the GO tree.

In both cases (GO and pathways terms), for each disease we identified the annotations significantly enriched within the corresponding seeds set (hypergeometric test, significance level 10^{-3}, with FDR Benjamini-Hochberg correction and annotation size greater than 10 and less than 300, to avoid bias [17]). The results are presented in Figures 5A and 5B where the median of the THs distributions for all 70 diseases at each iteration step, for both database and algorithm, are depicted. The better overall performances of DiaBLE *versus* DIAMOnD are clearly visible and their high significance is also reported (Wilcoxon rank sign test *P*-value less than 10^{-70}). In both cases (GO and pathways terms), for each disease we identified the annotations significantly enriched within the corresponding seeds set (hypergeometric test, significance level 10^{-3}, with FDR Benjamini-Hochberg correction and annotation size greater than 10 and less than 300, to avoid bias [17]). The results are presented in Figure 5A where the median of the THs distributions for all 70 diseases at each iteration step, for both database and algorithm, are depicted. The better overall performances of DiaBLE *versus* DIAMOnD are clearly visible and their high significance is also reported (Wilcoxon rank sign test *P*-value less than 10^{-70}). Figure 5B, computed as for the cross-validation case, shows that DiaBLE outperforms DIAMOnD starting from the initial 20 iterations.

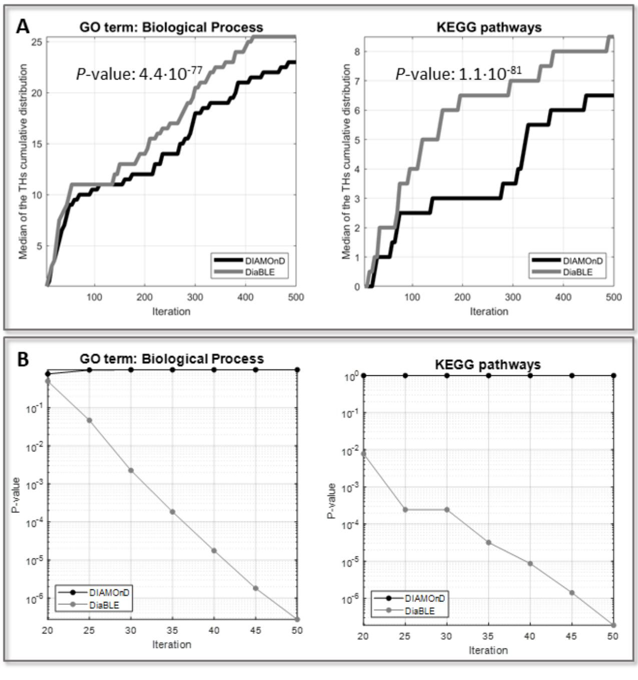


Fig.5. (**A**) Median of the cumulative True Hits (THs) distribution across the 70 diseases, for DiaBLE and DIAMOnD genes, for the cases of GO terms (left panel) and KEGG pathways (right panel). Wilcoxon signed rank test reveals in both cases a highly significant difference of the median values (*P-value* <10^{-70}) in favor of DiaBLE. (**B**) *P*-values corresponding to the initial 50 iterations. It is worth noting that increased overall performances are visible from the initial 20 iterations, thus confirming again the positive impact of the choice of an expanding gene universe in the top new disease genes.

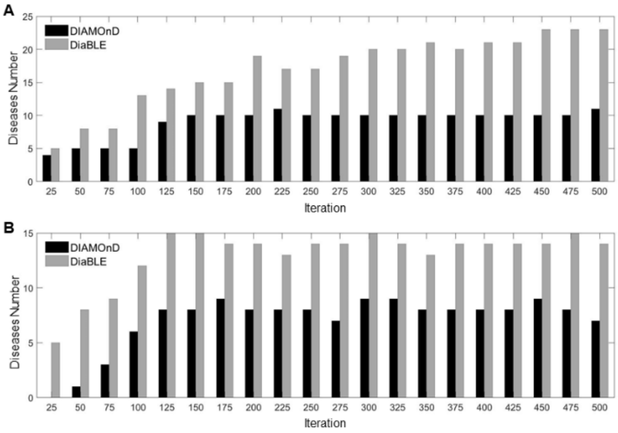


Fig. 6. Bar diagrams (panels **A**, **B** respectively for the cases of GO terms and KEGG pathways) showing the amount of diseases for which the performances are statistically better (Wilcoxon signed rank test, *P-value* < 10^{-3}) using DIAMOnD (black bars) and DiaBLE (grey bars).

As described for the cross-validation case, we studied differences in algorithms performances using a statistical comparison (Wilcoxon signed rank test, *P-value* < 10^{-3}) between DIAMOnD and DiaBLE by considering cumulative THs vectors for each disease. Again, for this analysis, we performed the comparison by taking into account different iteration values $N_{it} \in [25; 500]$: starting from the first 25 iterations until the final value, increasing their number with a step of 25 iterations. Figure 6 shows the obtained results for GO terms (panel *a*) and KEGG pathways (panel *b*): from the initial iterations (*i.e.* considering 25 iterations), DiaBLE outperforms DIAMOnD in a larger number of cases.

3.4 Enrichment analysis of top DIAMOnD and DiaBLE genes in two cancers

To provide also a biological analysis using DiaBLE genes, we considered two tumors which are not present in the original disease list [8]: the head and neck squamous cell carcinoma (HNSC) and the kidney renal clear cell carcinoma (KIRC).

Therefore, to obtain the associated seed sets (*i.e.* "known" disease genes), we used the database of disease-gene associations DisGeNET ([18], [19], www.disgenet.org/) with the following selection criteria: i) curated data, ii) $s > 0.2$ (where s is the DisGeNET score of disease-gene association level of evidence). We then performed an enrichment analysis in KEGG pathways for seed genes and for the top 50 genes identified by DIAMOnD and DiaBLE in order to recognize specific pathways in which the differences between DIAMOnD and DiaBLE are more evident (Supplementary Tables S2 and S3). We selected two tumors, HNSC and KIRC, because they produced somewhat "divergent results": in the first case, DIAMOnD and DiaBLE identified completely different lists of genes (3 out of 50 genes were equal); in the latter, the two lists were very similar (4 out of 50 were different) but with a very different ordering.

In HNSC, DiaBLE has been able to identify more statistically significant enriched pathways than DIAMOnD, just like those identified by seed genes. Some of these pathways, such as p53 and apoptosis, are characteristics of DiaBLE and the list of seed genes but they are not present in the DIAMOnD genes list. Moreover, according to the "Cancer Genome Atlas" data (TGCA, https://cancergenome.nih.gov/) for head and neck cancer, p53 is the most common genetic mutation with a population frequency of 72%. As opposed to the HNSC, in KIRC, we found almost the same genes in both DIAMOnD and DiaBLE lists and this explains the same enriched pathways obtained from KEGG analysis. Nevertheless, we found an intriguing difference in terms of statistically significance of enriched pathways exclusively in favor of DiaBLE, and it concerns several pathways involved in cancer (estrogen, PI3K-AKT, cGMP-PKG, TNF and AMPK signaling). Then, we looked at the position of genes in the two list (DiaBLE and DIAMOnD). We considered pathways in which the genes that "move-up" or "move-down" positions in the DiaBLE list, compared to DIAMOnD, are more frequently included. To this purpose, we looked at all pathway databases present in the GeneCard database (https://www.genecards.org/) (Figure 7). As shown in Figure 7, the large-scale differences in pathways enriched of prioritized genes concern Insulin, PI3K-AKT, AMPK, Notch and PKA in favor of DiaBLE algorithm ("move-up" genes) and NFkB, TGFb, HIF and Jak-Stat in favour of DIAMOnD algorithm ("move-down" genes). Most of the prioritized pathways from both algorithms are included among those enriched for seed genes, suggesting they could play a key role in the carcinogenesis of KIRC. Although it is difficult to speculate about the different biological relevance of all the identified pathways in KIRC pathogenesis, some of the pathways prioritized by DiaBLE, such as insulin, PDGF, PI3K-AKT and AMPK pathways can regulate mTOR signaling, whereas the remaining ones are involved in apoptosis and cell cycle. Interestingly, mTOR pathway is a primary target in the treatment of advanced KIRC, indicating its important role in this cancer [20]–[22]. Moreover, DiaBLE algorithm predicts, strongly than DIAMOnD, a role for Notch, Sonic-Hedgehog and PKA pathways in the pathogenesis of KIRC which cannot be evidenced using known disease genes. Notably, their potential role in this cancer is consistent with recent literature [23]–[25].

4 Conclusion

Inspired by the key insight about the importance of the connectivity significance for disease gene prediction provided by Ghiassan et al. [7], in this work we proposed a new version of the iterative process underlying the DIAMOnD algorithm. Since, at each algorithm iteration, nodes

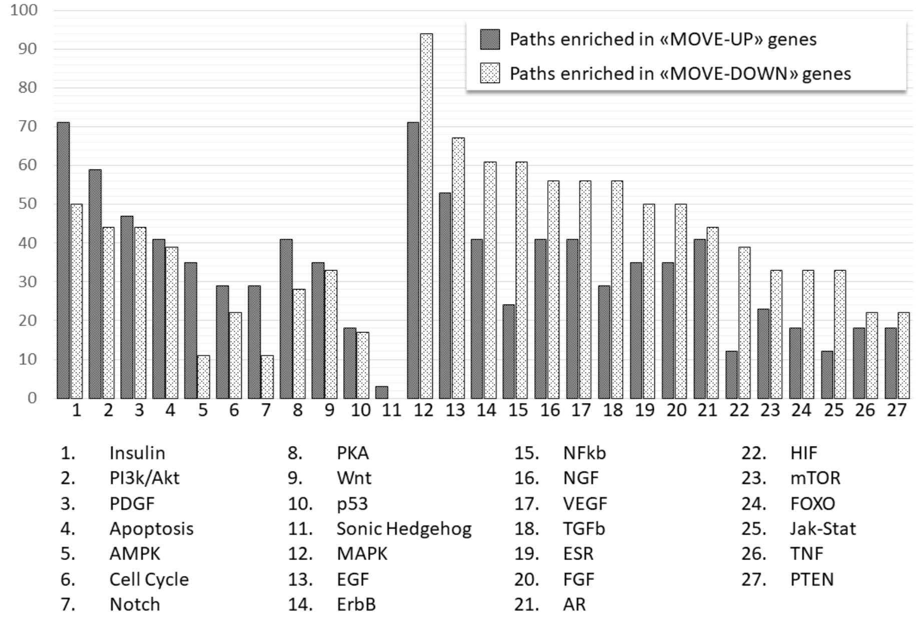


Fig. 7. Pathways including genes that "move-up" and "move-down" positions in DiaBLE compared to DIAMOnD.

that have no connection with the seed genes are not "competing" to become such, the DiaBLE algorithm considers as background model for the hypergeometric test the smallest local expansion of the current disease module: the resulting gene universe is thus composed of: i) seeds set ii) candidate genes (nodes having at least a link to the seeds set) and iii) first neighbors of the candidate genes. We showed the impact of such gene universe selection and how DiaBLE genes are related to a significant increase of general performances with respect to the DIAMOnD algorithm both for computational and biological validation. Moreover, on two specific cancers, we proved that both in the case of different predicted genes and in the case of very similar predicted genes but with different ordering – the DiaBLE algorithm provides more biological meaningful results compared to DIAMOnD. Finally, we note that comparative analysis among gene prioritization algorithms based on different strategies is beyond the scope of this study.

Acknowledgement

This work has been partially supported by a Sapienza University of Rome grant entitled "Network medicine-based machine learning and graph theory algorithms for precision oncology" - n. RM1181642AFA34C2.

References

[1] S. Y. Chan and J. Loscalzo, "The emerging paradigm of network medicine in the study of human disease," *Circ. Res.*, vol. 111, no. 3, pp. 359–374, Jul. 2012.

[2] M. Gustafsson *et al.*, "Modules, networks and systems medicine for understanding disease and aiding diagnosis," *Genome Med.*, vol. 6, no. 10, p. 82, 2014.

[3] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network Medicine: A Network-based Approach to Human Disease," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[4] J. Menche *et al.*, "Disease networks. Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, Feb. 2015.

[5] K. Ozturk, M. Dow, D. E. Carlin, R. Bejar, and H. Carter, "The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine," *J. Mol. Biol.*, vol. 430, no. 18 Pt A, pp. 2875–2899, Sep. 2018.

[6] Y. Bromberg, "Chapter 15: disease gene prioritization," *PLoS Comput. Biol.*, vol. 9, no. 4, p. e1002902, Apr. 2013.

[7] S. D. Ghiassian, J. Menche, and A.-L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome," *PLOS Comput. Biol.*, vol. 11, no. 4, p. e1004120, Apr. 2015.

[8] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nat. Rev. Genet.*, vol. 18, no. 9, pp. 551–562, 2017.

[9] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Am. J. Hum. Genet.*, vol. 82, no. 4, pp. 949–958, Apr. 2008.

[10] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinforma. Oxf. Engl.*, vol. 26, no. 8, pp. 1057–1063, Apr. 2010.

[11] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein–protein interactions," *J. Med. Genet.*, vol. 43, no. 8, pp. 691–698, Aug. 2006.

[12] S. Navlakha, M. C. Schatz, and C. Kingsford, "Revealing biological modules via graph summarization," *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, vol. 16, no. 2, pp. 253–264, Feb. 2009.

[13] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating Genes and Protein Complexes with Disease via Network Propagation," *PLOS Comput. Biol.*, vol. 6, no. 1, p. e1000641, gen 2010.

[14] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 523–536, Jul. 2012.

[15] S. Falcon and R. Gentleman, "Using GOstats to test gene lists for GO term association," *Bioinforma. Oxf. Engl.*, vol. 23, no. 2, pp. 257–258, Jan. 2007.

[16] S. Falcon and R. Gentleman, "Hypergeometric Testing Used for Gene Set Enrichment Analysis," in *Bioconductor Case Studies*, F. Hahne, W. Huber, R. Gentleman, and S. Falcon, Eds. New York, NY: Springer New York, 2008, pp. 207–220.

[17] D. Guala and E. L. L. Sonnhammer, "A large-scale benchmark of gene prioritization methods," *Sci. Rep.*, vol. 7, p. 46598, Apr. 2017.

[18] J. Piñero *et al.*, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D833–D839, Jan. 2017.

[19] J. Piñero *et al.*, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database J. Biol. Databases Curation*, vol. 2015, p. bav028, 2015.

[20] D. Fantus, N. M. Rogers, F. Grahammer, T. B. Huber, and A. W. Thomson, "Roles of mTOR complexes in the kidney: implications for renal disease and transplantation," *Nat. Rev. Nephrol.*, vol. 12, no. 10, pp. 587–609, 2016.

[21] D. Su, E. A. Singer, and R. Srinivasan, "Molecular pathways in renal cell carcinoma: recent advances in genetics and molecular biology," *Curr. Opin. Oncol.*, vol. 27, no. 3, pp. 217–223, May 2015.

[22] M. Ghidini *et al.*, "Clinical development of mTor inhibitors for renal cancer," *Expert Opin. Investig. Drugs*, vol. 26, no. 11, pp. 1229–1237, Nov. 2017.

[23] D. Jędroszka, M. Orzechowska, and A. K. Bednarek, "Predictive values of Notch signalling in renal carcinoma," *Arch. Med. Sci. AMS*, vol. 13, no. 6, pp. 1249–1254, Oct. 2017.

[24] V. Dormoy *et al.*, "The sonic hedgehog signaling pathway is reactivated in human renal cell carcinoma and plays orchestral role in tumor growth," *Mol. Cancer*, vol. 8, p. 123, Dec. 2009.

[25] B. Zhang *et al.*, "G Protein Alpha S Subunit Promotes Cell Proliferation of Renal Cell Carcinoma with Involvement of Protein Kinase A Signaling," *DNA Cell Biol.*, vol. 36, no. 3, pp. 237–242, Mar. 2017.