2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

Network-based disease gene prioritization by hitting time analysis

Leandro de A. Lima*, Sérgio N. Simões*, Ronaldo F. Hashimoto,
David C. Martins-Jr, Helena Brentani, Guilherme O. Mota

***Abstract*—Many methods have been published to prioritize genes using network theory. By using protein-protein interaction (PPI) data, it is possible to use mathematical features to rank and prioritize genes products in the network. Taking into account that genes related to the same diseases tend to connect, in the network structure, the prioritization methods search for candidate genes in the neighborhood of other genes already known to be related to a specific phenotype. Unfortunately, some existing algorithms can not deal well with highly connected genes, and some of them end up being related by chance to the disease being studied. We propose a pure method, with no need of adjustments, based on the hitting time of a random walk in PPI networks. This method captures information of the whole network and can equally prioritize genes regardless of its degree. We tested the efficiency of our method prioritizing candidate genes for Attention-Deficit/Hyperactivity Disorder (ADHD). The proposed method was able to give a good rank to genes that have genetic association with ADHD and was able to prioritize a large proportion of genes prioritized by other random-walk-based methods.**

***Index Terms*—hitting time, gene prioritization, PPI networks.**

I. Introduction

In the last 10 years, many studies have been published about finding the missing heritability in complex diseases [1]–[3]. These disorders, unlike the Mendelian ones, can be caused by many factors, having genetic and environment influence [1], [4]. Finding the causes of such disorders has been a problem. With the advent of high-throughput methods to obtain molecular data, it has become possible to use different types of data to improve the search for the causes of complex disorders [1], [7]. At present, there are many public databases with information about genes, and its products, the proteins, as well as information of how they interact with each other. Complex diseases can be caused by different variants (common, rare, *de novo*) acting together in different genes, and it is also accepted that if one gene have more than 4 variants in different trios it can be considered pathogenic.

The study of the interaction between proteins, that are products of genes, has been confirmed as a promising way to find out new candidate genes or biological functions as being related to a disease of interest [7], [9], [10].

Protein-protein interaction (PPI) networks can be used in different ways to find genes related to a disease. Usually such data can help to find new candidate genes or to validate findings from genome-wide association studies (GWAS), exome analysis and expression analysis. Genes already associated to a disease could be used as "controls" compared to a list of candidates ranked by a new method in two ways: (1) we can observe how many "control" genes are present in the new candidate gene list, because it is expected that some genes already described will be well ranked by the method, and (2) we can observe if genes in the candidate gene list are "near" or even directly connected to the "control" genes, assuming that genes related to a disorder or same pathway must be more connected. Some of network methods are based only on local information (e.g. direct neighbours and shortest paths) and others are based on global information, by considering the whole topology of PPI networks (e.g. random walks and network propagation). Currently, global methods [9], [10], [18] are commonly used to prioritize genes and investigate a neighbourhood of genes that are known to be related to a disease ('seeds').

In this context, Kohler [9] proposed to use a random walks with restart (RWR) method to prioritize candidate disease genes. It takes as input a PPI network, a set of seeds and candidate genes and outputs the set of candidates ranked by a score. The RWR based association score of a given gene can be computed iteratively with the formula $p^{t+1} = (1-r)Wp^t + rp^0$. Here, W is the adjacency matrix (column-normalized) of the network. The vector p^t holds in its i-th element the probability of the random walk be on node i at time step t. Furthermore, p^0 is the initial probability vector and r is the probability of restarting the random walk at a given node.

Unfortunately, PPI databases have a bias in their degree distribution due to the manner in which different proteins are assayed for interactions. Some genes products are much more studied than others and consequently have more connections in the PPI networks.

Aiming to circumvent this problem, a suite of statistical adjustment methods (called DADA [6]) was proposed in order to correct this bias in global methods of disease gene

* These authors contributed equally to this work.
L.A. Lima is with Inter-institutional Grad Program on Bioinformatics, University of São Paulo, SP, Brazil (e-mail: llima@ime.usp.br)
S.N. Simões is with Federal Institute of Espírito Santo, Serra, ES, Brazil (e-mail: sergions@ime.usp.br)
R.F. Hashimoto is with Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil (e-mail: ronaldo@ime.usp.br)
D.C. Martins-Jr is with Center for Mathematics, Computation and Cognition, Federal University of ABC, Santo André, SP, Brazil (e-mail: david.martins@ufabc.edu.br)
H. Brentani is with Institute of Psychiatry, University of São Paulo, São Paulo, SP, Brazil (e-mail: helena.brentani@gmail.com)
G.O. Mota is with Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil (e-mail: mota@ime.usp.br)
Manuscript received July 31, 2014; revised September 8, 2014.

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.22

prioritization. Whilst RWR detects only higher degree genes in the seeds neighbourhood, DADA intends to detect both high and low degree related genes. They proposed to accomplish this by integrating one of three different statistical adjustments with one of three different uniform prioritization methods.

In this work, we apply an established network-based analytical algorithm to find genes related to a complex disorder using the whole PPI network to prioritize genes, without being biased by highly connected genes. In this approach, all possible paths between pairs of genes are considered, providing a rich amount of information.

We choose the Attention-Deficit/Hyperactivity Disorder (ADHD) as our case study of complex disorder since GWAS significant results were not achieved yet and it seems clear that an integrative approach could contribute to its understanding [20], [21]. A recent paper was published by a Psychiatric GWAS Consortium using a PPI network approach, but the authors used a limited version of the local method that adds to the network only genes that share the same seeds as neighbors [19]. The genetics of ADHD, like most of psychiatric disorders, is complex due to both a non-Mendelian inheritance pattern and a huge heterogeneity in phenotype, suggesting the relevant role of multiple genes in the etiology of these disorders [5]. We used a list of the ADHD database [8] as our gold standard list to confirm genes prioritizided by application of hitting time.

II. Background

A. Hitting time method

In what follows, let G be a PPI network, let $S = \{s_1, \ldots, s_k\}$ be a set of seeds, i.e., a set of nodes that are previously known to be related to the disease that will be investigated. Finally, let C be a set of candidates, i.e., the set of nodes we want to give a rank, accordingly to their relation with the disease. One can ask the following question: how many steps we expect that a random walk starting at c takes to reach s for the first time? The *hitting time* from c to s, which we denote by $H(c, s)$, is defined as this expected time. Note that in order to calculate the hitting time from c to s, all paths between them have to be considered. Thus, information contained in the whole network is encapsulated in the hitting time.

For each candidate $c \in C$, we calculate the hitting time from c to s_i, for all seeds $s_i \in S$. Then, the *HT-average* of the candidate c, denoted by $HT_{\text{avg}(c)}$, is defined as the average of all these values, i.e.,

$$HT_{\text{avg}}(c) = \sum_{i=1}^{k} \frac{H(c, s_i)}{k}. \quad (1)$$

The smaller this value the better is the rank of c. Thus, the best ranked gene is that one with smallest HT-average. We remark that, instead of the average, we could use the maximum or minimum among the hitting time values from a candidate to all seeds, but the difference would be almost irrelevant.

In our method, for a candidate to receive a good rank (small HT-average), it is not enough to have high degree on the PPI network or being close to the seeds. Imagine two scenarios in which genes with different degree should or not be prioritized:

(i) Suppose that a candidate has small degree, but all its neighbours either are seeds or are nodes that are connected to all seeds. In this case, probably the candidate is related to the disease we are considering. Thus, it would be interesting that the hitting time method give a good rank to such candidate.

(ii) Suppose that a candidate has high degree, but none of them are seeds and all its neighbours are very far from the seeds. We expect that a good method gives a bad rank to such candidate.

The hitting time method is rigorous, in the sense that in order to obtain a good rank, the candidate must have many good "connections" to the seeds.

III. Methods

A. PPI network dataset

We constructed a network using genes mapped in protein-protein interaction (PPI) databases and their physical interactions to analyze the mathematical properties that can give information about the network components. We used three important human PPI databases: Human Protein Reference Database - HPRD [14], The Molecular Interaction Database - MINT [16] and The IntAct Molecular Interaction Database [15]. The databases were downloaded and integrated using official gene symbol and Entrez ID. With the union of the three databases, the whole human interactome constructed has 14276 nodes (proteins/genes) and 88525 physical interactions, supported by at least one piece of direct experimental evidence demonstrating physical interaction between two human proteins in at least one database. We use the network composed by the largest connected component of the above network, which contains 12373 proteins and 74337 interactions.

B. ADHD GWAS dataset

To choose the seeds for this analysis in ADHD, we selected the 50 resulting SNPs with P-value $\leq 10^{-5}$ from Neale et al. [11] meta-analysis, mapped in 14 genes. This meta-analysis included only children/adolescents and studies that used the same categorical diagnostic variables, decreasing chances of phenotypic heterogeneity. Taking into account this list of 50 SNPs, we searched in a 100kb upstream and downstream interval, SNPs in linkage disequilibrium (LD) for $R^2 \geq 0.8$ with the SNPs in the list, using the International HapMap Project (available at http://hapmap.ncbi.nlm.nih.gov/) data phase III/Rel#2, CEU population samples. After using the LD data, we obtained a set of 30 genes, but only 19 were mapped in the PPI databases. We used this set as seeds to perform the network analysis in this work. Similar to the method performed by Poelmans et al. [20], we use GWAS results to construct and analyze a protein-protein network.

C. Mathematical details of the Hitting Time Method

Let G be a network with n nodes (say, $1, \ldots, n$). Let A_G be the *adjacency matrix* of the graph G, i.e., an $n \times n$ matrix

such that, for every $1 \leq i < j \leq n$, we have $A_G(i,j) = 1$ if i is connected to j in G. Furthermore, we put $A_G(i,j) = 0$ in the case i and j are not connected. This matrix represents the whole graph.

Given an $n \times n$ matrix A, we say that λ is an *eigenvalue* of A if there exists a non-null vector v such that $Av = \lambda v$. In this case, we say that v is an *eigenvector* of λ.

Given a node v of the network G, a random walk starting at a node w is a sequence of random steps such that the following holds: we start at w and we move to a neighbour z of w with probability $1/d_G(w)$, where $d_G(w)$ denotes the number of neighbours of w in the network G. Now we apply another step, looking to the neighbours of z and so on. Formally, let $p_{ij} = 1/d_G(i)$ if i and j are neighbours, and let $p_{ij} = 0$ otherwise. Let $M = (p_{ij})_{1 \leq i < j \leq n}$ be the *random walk matrix* of G. The random walk is completely defined by the matrix M, because if the random walk is at a vertex i, one can just look to the i-th row of M to calculate the probability of being in a vertex v in the next step.

Symmetric matrices have many properties that can help us to obtain information about networks. Unfortunately, the random walk matrix M is not symmetric unless G is regular, i.e., all nodes have the same number of neighbours (usually that is not the case for PPI networks). But it is easy to check that the matrix M can be seen as the matrix obtained by the product of the matrices D and A, i.e., $M = DA$, where D is the diagonal $n \times n$ matrix such that $D_{ii} = 1/d_G(i)$. Now, after some calculations, one can check that the matrix $N = D^{-1/2}MD^{1/2} = D^{1/2}AD^{1/2}$ is symmetric. Since N is a symmetric matrix, it has n real eigenvalues and, furthermore, it is well known that N can be written as a sum of products of eigenvalues and eigenvectors. Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the (real) eigenvalues of N and let $v_1, \ldots, v_n$ be the corresponding eigenvectors of unit length. We have

$$N = \sum_{i=1}^{n} \lambda_i v_i v_i^T,$$

where v_i^T is the transpose of v_i.

Let v and w be nodes of G and consider a random walk on G starting at v. We denote by $X_{v,w}$ the random variable that denotes the number of steps that a random walk takes to leave v and reach w for the first time. The *hitting time from v to w* of a random walk in G is defined as follows.

$$H(v,w) = \mathbb{E}(X_{v,w}).$$

In other words, $H(v,w)$ is the expected time of the number of steps that a random walk starting at v takes to reach w for the first time. Thus, in order to calculate the hitting time, all the possible paths between v and w have to be considered, meaning that connections of the whole network have to be taken into account. Then, we applied the Lovász equation [18] that allows to calculate the hitting time between all node pairs.

Theorem 1. Let G be a connected network with n vertices (say $1, \ldots, n$) and m edges. Let v and w be nodes of G. Then,

$$H(v,w) = 2m \sum_{i=2}^{n} \frac{1}{1-\lambda_i} \left(\frac{v_i(w)^2}{d_G(w)} - \frac{v_i(v)v_i(w)}{\sqrt{d_G(v)d_G(w)}} \right),$$

where $v_i(x)$ denotes the x-th coordinate of the vector v_i.

Let G be a graph and let $\mathcal{V}$ be a basis of eigenvectors for the matrix N associated with G. Unfortunately, for many graphs, the associated matrix N has eigenvalues with large multiplicity, which implies that $\mathcal{V}$ may not be orthogonal, as required in Theorem 1. To find the orthonormal basis of eigenvectors, we implemented the "Gram-Schmidt process" to obtain an orthogonal basis from $\mathcal{V}$.

IV. Results and discussion

We make use of the hitting time on the PPI network in order to prioritize genes potentially associated with complex diseases.

The main purpose of the hitting time method is to avoid the bias of the degree in a network, generated by nodes which are highly connected just because they are more often investigated in the literature. We tested the efficiency of the method by applying it to prioritize genes that are related to Attention-Deficit/Hyperactivity Disorder (ADHD). The PPI network considered here is composed by the largest connected component of the union of three datasets, namely, HPRD, MINT and IntAct.

In order to check if the first genes prioritized by HT have some relation with ADHD, we searched on ADHDgene website the genes ranked in the first ten positions by HT, and six of them showed a genetic relation to ADHD (see Table I).

We remark that the genes with the best ranks are in genomic regions known to be related to ADHD, either because of single nucleotide polymorphisms (SNPs) or copy number variations (CNVs), and yet some were already reported in literature as having a significant genetic relation.

Furthermore, the bias generated by the highly connected nodes does not affect the hitting time method. In spite of the fact that the PPI network used here contains more than 1800 nodes that have at least 20 interactions, and more than 150 nodes that have at least 100 interactions, in the top 20 ranked by hitting time displayed maximum degree of 36 and average degree of 7.3 links. See Table I for more information (average distance to the seeds and average hitting time to the seeds) about the nodes that receive the 10 best ranks (the smallest average values of the hitting time from the candidates to the seeds).

We also compared the efficiency of our method with two other methods: (i) the first method is based on a random walk with restart called here by RWR; (ii) the second, called DADA, can be seen in [6]. We applied the methods RWR and DADA to the same PPI network to compare their results.

Fig. 1 contains the seeds (genes with red border) used in the analysis and the top 10 genes ranked by each method. The top 10 genes ranked by DADA are in yellow and the top 10 genes ranked by Hitting Time are in blue. The green genes are among the top 10 ranked by both methods. The genes that are reported by ADHDgene website as being related to ADHD are with diamond shape.

In order to compare how HT and DADA behave similarly in long term, we plotted the overlap of genes prioritized by both methods. For the first n genes prioritized, n ranging

TABLE I
TABLE OF GENES RANKED IN THE TEN FIRST POSITIONS BY HT (COLUMN "AVERAGE HITTING TIME") AFTER REMOVING SEEDS OF THE RESULTS. THE RANKS GIVEN BY DADA ARE IN COLUMN "DADA RANK". COLUMN "AVERAGE DISTANCE" SHOWS THE AVERAGE DISTANCE FROM THE CORRESPONDING GENE TO THE SEEDS. THE LAST COLUMN SHOWS GENETIC ASSOCIATIONS OF THE CORRESPONDING GENES TO ADHD.

Gene	HT rank	DADA rank	Average distance	Average hitting time	Genetic relation with ADHD reported by ADHDgene
FADS1	1	1	4.15	33062.09	Significant for gene-based test [25]
AK1	2	2	3.55	33364.79	Indirect association
PTGS1	3	3	3.95	33434.11	Mapped by significant region (9q33) [22]
MYO15B	4	4	3.40	33473.18	Mapped by CNV [21]
PCID2	5	7	3.30	33549.58	No reported association
GMFG	6	10	4.05	33592.15	No reported association
PPP1R16A	7	9	3.35	33607.78	No reported association
PHP14	7	25	4.35	33608.78	No reported association
CUTC	9	8	3.25	33618.43	No reported association
SORCS1	10	16	4.05	33623.17	Literature-origin [23]
ZP4	10	17	4.05	33623.17	Mapped by CNV [24]
PRB4	10	15	4.05	33623.17	No reported association

from 1 to 12353, the overlap is shown in Figure 2. Until the 2000 first genes, the plot shows that HT was able to prioritize approximately 80% of genes prioritized by DADA. However, Figure 3 shows that DADA tends to prioritize more connected nodes, whilst HT was able to prioritize genes with low degree.

V. CONCLUSION

The efficiency of the Hitting Time method proposed in this paper was tested by prioritizing candidate genes for Attention-Deficit/Hyperactivity Disorder (ADHD). The method was able to retrieve genes that have genetic association to ADHD. To test the efficiency of the method we also compared the results with results obtained by other two well known methods for genes prioritization (one based on random walk with restart (RWR) and the other, called DADA, composed by a mix of strategies). The hitting time method was able to give similar ranks to the 20 first genes ranked by DADA (Table I).

We also showed that HT had a consistent behavior and was not affected by the bias of highly connected vertices. Besides HT was able to prioritize a large proportion of genes prioritized by DADA, no manual adjustment was necessary in order to choose more connected or less connected nodes to be prioritized.

We remark that hitting time has been successfully applied as an inference method to measure the network modularity, and in particular to suggest the subtypes of diseases based on a heterogeneous network [17]. In this paper, Yao et al. developed a Hitting-Time-based method, called CIPHER-HIT, for assessing the modularity of disease gene predictions and credibly prioritizing disease-causing genes.

ACKNOWLEDGMENT

We thank IME–USP for the pleasant working environment provided at the CCSL building. Such an environment made possible the integration between different research groups, giving rise to this work.

Leandro de Araújo Lima was supported by CAPES (10682/13-9), Sérgio Nery Simões was supported by Instituto Federal do Espírito Santo, David C. Martins-Jr was supported by CNPq (559955/2010-3), Ronaldo F. Hashimoto was supported by CNPq (308011/2013-9) and FAPESP (2010/52138-8), Helena Brentani was supported by FAPESP (2011/06594-4) and Guilherme O. Mota was supported by FAPESP (2013/11431-2). Finally we thank FAPESP grant 2011/50761-2, CNPq, CAPES, NAP eScience - PRP - USP.

REFERENCES

[1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH,

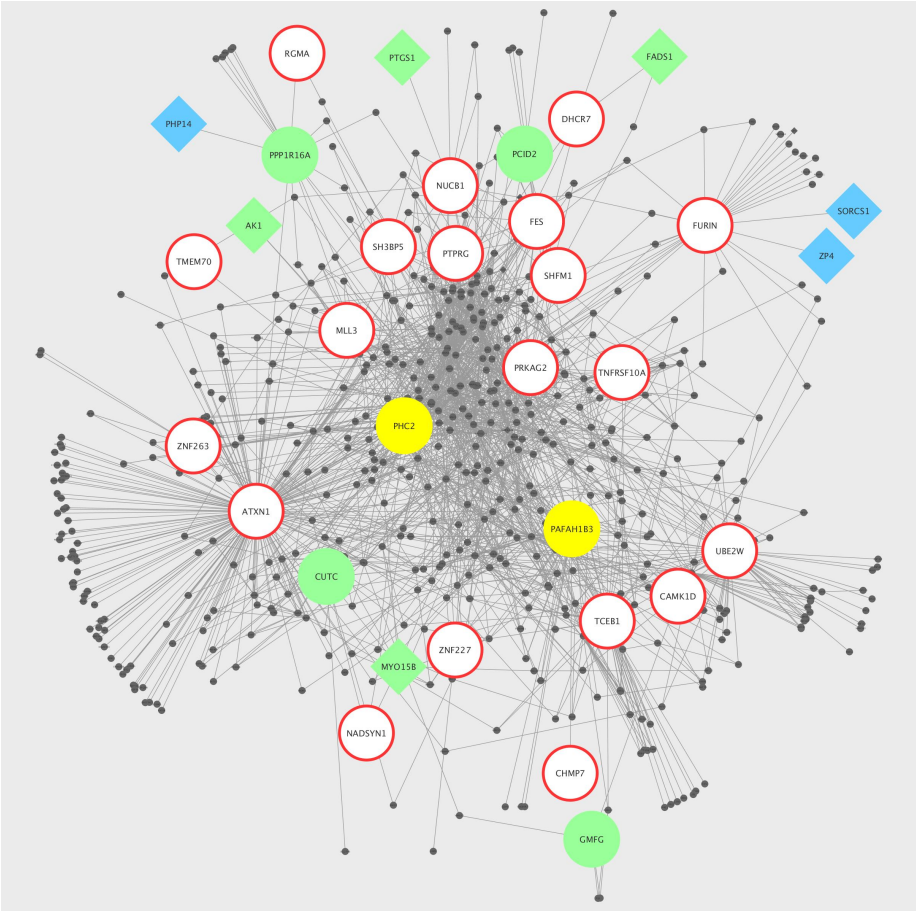

Fig. 1. Network with genes well ranked by Hitting Time (HT) and DADA. The genes with red border are the seeds. The genes in yellow are in top 10 ranked genes by DADA, the blue genes are in top 10 ranked genes by HT and genes in green are in top 10 ranked by both methods. The genes with diamond shape have significant genetic relation to ADHD reported by ADHDgene website.

Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53.

[2] Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010 Jun;11(6):446-50.

[3] Koch L. Disease genetics: Insights into missing heritability. Nat Rev Genet. 2014 Apr;15(4):218.

[4] Akutagava-Martins GC, Salatino-Oliveira A, Kieling CC, Rohde LA, Hutz MH. Genetics of attention-deficit/hyperactivity disorder: current findings and future directions. Expert Rev Neurother. 2013 Apr;13(4):435-45.

[5] Hamshere, ML et al. (2013). High loading of polygenic risk for ADHD in children with comorbid aggression. Am J Psychiatry, 170, 8:909-16.

[6] Erten, S, Bebek, G, Ewing, RM, Koyutrk, M (2011). DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. BioData Min, 4:19.

[7] Barabsi, AL, Gulbahce, N, Loscalzo, J (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet., 12, 1:56-68.

[8] Zhang, L., S. Chang, Z. Li, K. Zhang, Y. Du, J. Ott and J. Wang (2012). ADHDgene: a genetic database for attention deficit hyperactivity disorder. Nucleic Acids Res 40(Database issue): D1003-1009.

[9] Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics 82: 949958.

[10] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS computational biology 6: e1000641.

[11] Neale, B.M. et al (2010) Meta-analysis of genome-wide association studies of attention-decit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 49(9), 884-897.

[12] Larsson H, Chang Z, D'Onofrio BM, Lichtenstein P. The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan. Psychol Med. 2013 Oct 10:1-7.

[13] Faraone SV, Perlis RH, Doyle AE et al. Molecular genetics of attention-deficit/hyperactivity disorder. Biol. Psychiatry57(11),13131323(2005).

[14] Keshava Prasad, TS, Goel, R, Kandasamy, K, Keerthikumar, S, Kumar, S, Mathivanan, S, Telikicherla, D, Raju, R, Shafreen, B, Venugopal, A, Balakrishnan, L, Marimuthu, A, Banerjee, S, Somanathan, DS, Sebastian, A, Rani, S, Ray, S, Harrys Kishore, CJ, Kanth, S, Ahmed, M, Kashyap, MK, Mohmood, R, Ramachandra, YL, Krishna, V, Rahiman, BA, Mohan, S, Ranganathan, P, Ramabadran, S, Chaerkady, R, Pandey, A (2009). Human Protein Reference Database–2009 update. Nucleic Acids Res., 37, Database issue:D767-72.

[15] Kerrien, S, Aranda, B, Breuza, L, Bridge, A, Broackes-Carter, F, Chen, C, Duesbury, M, Dumousseau, M, Feuermann, M, Hinz, U, Jandrasits, C, Jimenez, RC, Khadake, J, Mahadevan, U, Masson, P, Pedruzzi, I, Pfeiffenberger, E, Porras, P, Raghunath, A, Roechert, B, Orchard, S, Hermjakob, H (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res., 40, Database issue:D841-6.

[16] Licata, L, Briganti, L, Peluso, D, Perfetto, L, Iannuccelli, M, Galeota, E, Sacco, F, Palma, A, Nardozza, AP, Santonico, E, Castagnoli, L, Cesareni, G (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res., 40, Database issue:D857-61.

[17] Yao, X, Hao, H, Li, Y, Li, S (2011). Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network. BMC Syst Biol, 5:79.

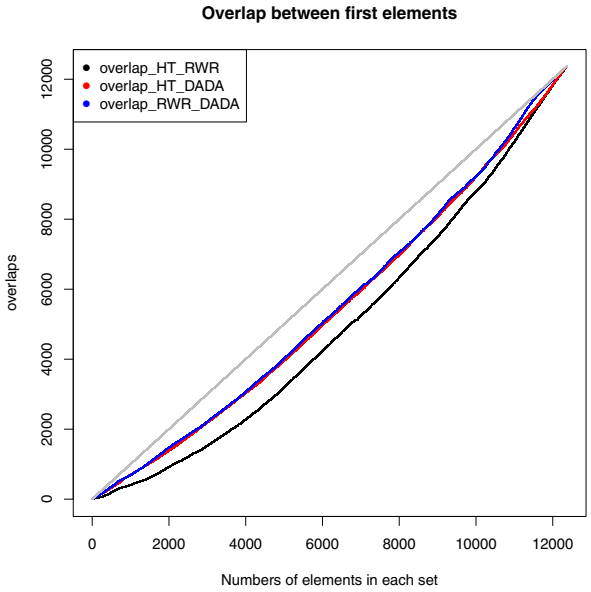

Fig. 2. Comparison between Hitting Time (HT), DADA and Random-walk with Restart (RWR), showing the overlaps between the methods, taken in pairs. The lines show how many genes are prioritized simutaneously by each pair of methods. The plot shows that HT can prioritize a high proportion of genes prioritized by DADA, as DADA does when compared to RWR. On the other hand, the overlap between HT and RWR lists is remarkedly smaller. It is important to notice that the gray line shows the superior limits for overlaps between two lists (100%), which indicates that many genes prioritized by one method are not prioritized by the other method.

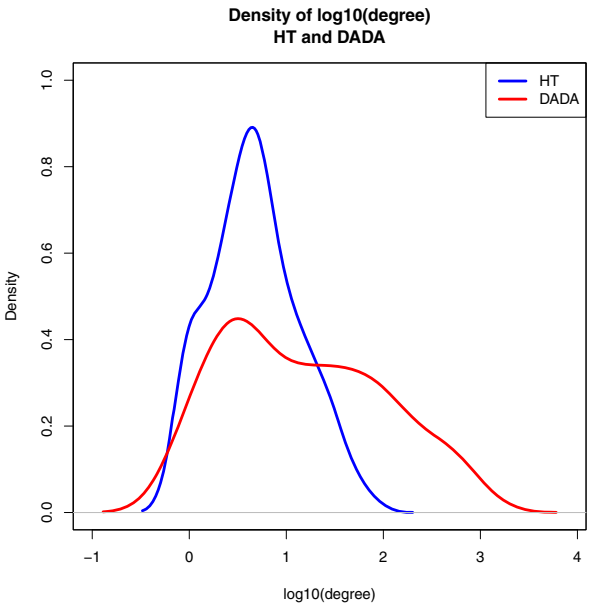

Fig. 3. Plot showing the degree (log10) density of 100 first genes prioritized by HT and DADA. Although many genes are the same, it is possible to see that DADA tends to prioritize more connected nodes than HT.

[18] Lovász, L. (1993) Random walks on graphs: a survey. Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993), 353:397.

[19] Yang, L, Neale, BM, et al. (2013). Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. Am. J. Med. Genet. B Neuropsychiatr. Genet. 162B(5):419-30

[20] Poelmans, G, Pauls, DL, Buitelaar, JK, Franke, B (2011). Integrated genome-wide association study findings: identification of a neurodevelopmental network for attention deficit hyperactivity disorder. Am J Psychiatry. 168(4):365-77

[21] Elia, J, Glessner, JT et al. (2012). Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder, Nat Genet. 44(1):78-84.

[22] Romanos M., Freitag C., Jacob C., Craig D. W., Dempfle A., Nguyen T. T., Halperin R., Walitza S., Renner T. J., Seitz C., Romanos J., Palmason H., Reif A., Heine M., Windemuth-Kieselbach C., Vogler C., Sigmund J., Warnke A., Schafer H., Meyer J., Stephan D. A. and Lesch K. P. (2008) "Genome-wide linkage analysis of ADHD using high-density SNP arrays: novel loci at 5q13.1 and 14q12." Mol Psychiatry, 13(5): 522-30.

[23] Lionel A. C., Crosbie J., Barbosa N., Goodale T., Thiruvahindrapuram B., Rickaby J., Gazzellone M., Carson A. R., Howe J. L., Wang Z., Wei J., Stewart A. F., Roberts R., McPherson R., Fiebig A., Franke A., Schreiber S., Zwaigenbaum L., Fernandez B. A., Roberts W., Arnold P. D., Szatmari P., Marshall C. R., Schachar R. and Scherer S. W. (2011) "Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD." Sci Transl Med, 3(95): 95ra75.

[24] Williams N. M., Zaharieva I., Martin A., Langley K., Mantripragada K., Fossdal R., Stefansson H., Stefansson K., Magnusson P., Gudmundsson O. O., Gustafsson O., Holmans P., Owen M. J., O'Donovan M. and Thapar A. (2010) "Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis." Lancet, 376(9750): 1401-8.

[25] Mick E., Todorov A., Smalley S., Hu X., Loo S., Todd R. D., Biederman J., Byrne D., Dechairo B., Guiney A., McCracken J., McGough J., Nelson S. F., Reiersen A. M., Wilens T. E., Wozniak J., Neale B. M. and Faraone S. V. (2010) "Family-based genome-wide association scan of attention-deficit/hyperactivity disorder." J Am Acad Child Adolesc Psychiatry, 49(9): 898-905 e3.