Reproducing and updating results from *Uncovering disease-disease relationships through the incomplete interactome*

Robin Petit[1] and Tom Lenaerts[1]

[1]Université Libre de Bruxelles

Abstract

This paper intends to work on the results exposed in Menche et al. (2015) in order to reproduce them, and update the different datasets to check if the presented procedure, which intends to be a systematic analysis, still stands for current version of the interactome, and if the results exposed have shifted towards a more or less significant level.

1 Introduction

An interactome is a graph containing all the molecular interactions within a cell. This notion appeared in Sanchez et al. (1999) (for the drosophila) and the need to thoroughly study this structure has been expressed by Barabasi and Oltvai (2004).

The interactome presented in Menche et al. (2015) has been used because authors showed that it is sufficiently complete to be systematically studied. Even though the main idea was to study the diseases behaviour inside the graph, different studies can be applied, e.g. drug-disease correlation Yu et al. (2016), or digenic diseases Gazzo et al. (2015).

In Menche et al. (2015), authors applied disease genes databases (in particular OMIM and GWAS) on the human interactome in order to determine the properties of their distribution in the graph. Major results were that: firstly diseases tend to *cluster* in denser subgraphs than the interactome itself (shown by bigger largest connected component than expected in random interactome subgraphs), secondly that phenotypically close diseases tend to overlap on a significant amount of genes.

Authors of the original paper also remind that the interactome is incomplete and estimated around 20%-complete for the interactions by the estimation that the interactome contains roughly 6.5×10^5 interactions in Stumpf et al. (2008), and around 54%-complete for the proteins involved by the estimations that the interactome contains roughly 2.5×10^4 nodes in Amaral (2008).

NOTE: references expressed as Sx refer to the original paper's supplementary materials. Any other reference is to this very paper, unless explicitly mentioned.

2 Reproducing results

The first part of this paper focuses on the reproduction of exposed results in Menche et al. (2015), namely the disease modules propensity to cluster into highly connected components and the significantly lower separation indicator values for highly gene related disease pairs.

The interactome used in Menche et al. (2015) contains 13460 genes and 141296 physical links constructed on several interactions databases including BioGRID, IntAct, TRANSFAC, MINT, HPRD, KEGG, BIGG, CORUM and PhosphitePlus and is available to download as supplementary material.

Diseases from OMIM (Online Mandelian Inheritance in Man) and GWAS (Genome-Wide Association Studies) related to at least 20 genes have been compiled to be applied to the interactome.

2.1 Clustering of disease modules

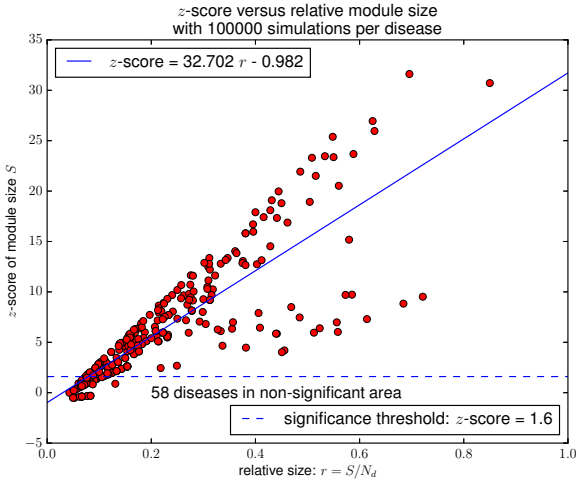

Figure 1: z-score of largest connected component size vs relative module size.

Figure S4.b plots the relative size of each disease module versus its relative size (defined as the quotient of the largest

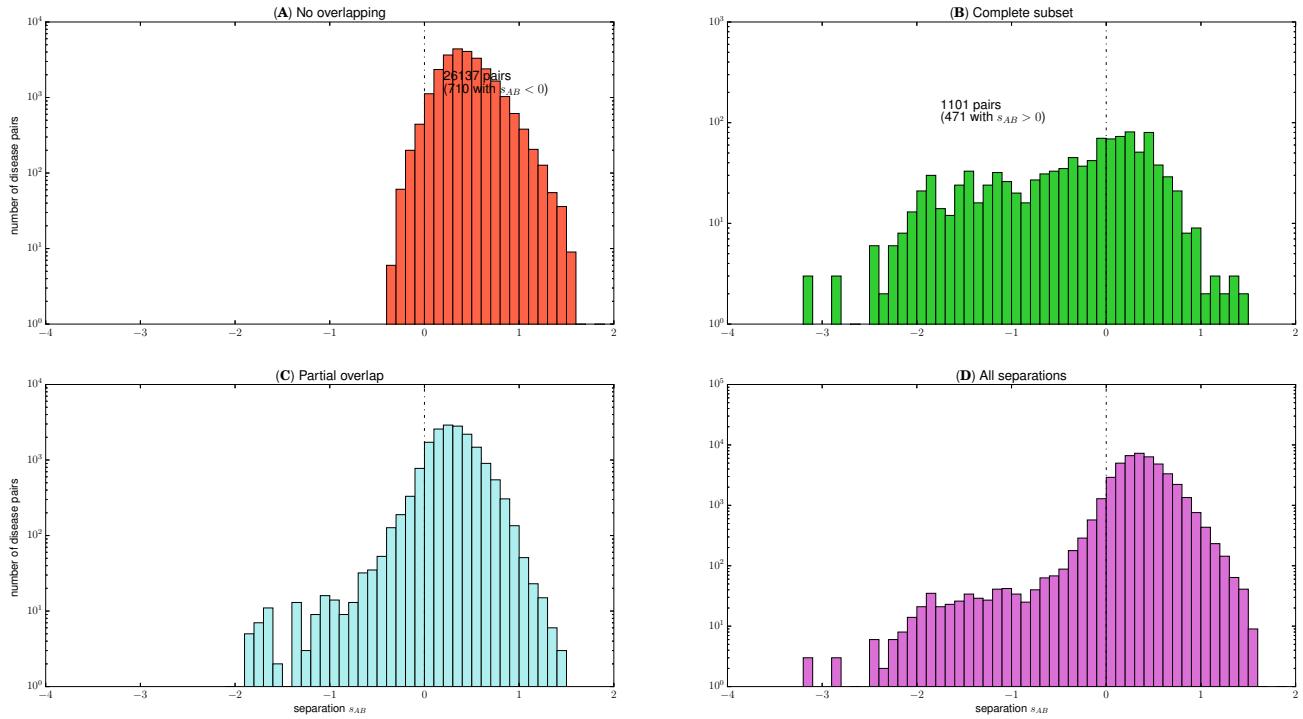

Figure 2: Disease pairs separation.

connected component size by the number of genes related to the disease).

When plotting the same data making 10^5 random simulations per disease and setting the significance threshold to be 1.6[1], the obtained result is shown on Figure 1, which fits the one presented in the original paper.

This figure shows that several diseases do not present a significantly larger largest connected component than expected by chance, but that these diseases have a thin relative size (< 0.2, so that less than 20% of their related genes are not present in the current interactome). Also, diseases with bigger relative size have a higher z-score, leading to think that a more complete interactome, with higher coverage of the diseases would increase the significance of the result.

2.2 Separation distribution

Original paper's figure 3.K-L plots the separation distribution of the disease pairs according to their overlapping score (C-score and J-score defined respectively as $|A \cap B| \,/\, \min(|A|, |B|)$ and $|A \cap B| \,/\, |A \cup B|$ for A and B two diseases). Disease pairs AB having a null J-score (and thus C-score) do not share any gene since $|A \cap B| = 0$. Disease pairs AB having a C-score equal to 1 are either identical (if their J-score equals 1) or one is a strict subset of the other.

Figure 2 plots the distribution of s_{AB} separation indicator. Distribution shown on (**A**) and (**B**) fit almost exactly the one presented in the article, the only difference being that for non-overlapping disease pairs, the amount of pairs having a separation value below 0 is 710 versus 717 in the original paper, being totally non-significant since 7 pairs represent less than 0.03% of all the non-overlapping pairs set.

3 Databases update

3.1 Interactome update

In order to update the interactome, the tool *inter-tool* presented in Catabia et al. (2017) has been used. Latest datasets from BioGRID, IntACT, and the Database of Interacting Proteins (DIP) (on July 25th) and of MINT (on July 27th) have been downloaded and merged via inter-tools.

Inter-build (one of the programs from Inter-tools) requires datasets in the PSI-MITAB format, an extension of the PSI-MI format, described in Kerrien et al. (2007), and outputs a tsv file described in Catabia et al. (2017).

The outputted file by Inter-build and the interactome provided with the original paper both use Entrez gene IDs, they can therefore easily be merged. In order to merge these, the script `merger.py` has been written. As these two tsv files have a different format (different columns), only the gene

[1] Considering the distribution to be normal as in Barraez et al. (2000), a z-score ≥ 1.6 represents a p-value ≤ 0.05 which corresponds to considered *significant* results.

IDs are outputted by `merger.py`.

This newer version merged with the original interactome yield a new graph having 17786 nodes and 370326 edges (so a bit more than 1.3 times the initial amount of nodes, and more than 2.6 times the initial amount of edges).

This newer version is around 71%-complete for the proteins involved and 57%-complete for the interactions according to estimations from Stumpf et al. (2008) and Amaral (2008).

3.2 Disease genes update

Disease genes considered in this update are the same 299 diseases studied in Menche et al. (2015). An update of these is yet to be done in further investigations.

3.3 Updated results

By defining the density of a graph $\Gamma = (V, E)$ as being $d(\Gamma) := |E| / \binom{|V|}{2}$, we find that the newer interactome is more than 1.5 times denser than the one used in Menche et al. (2015) (0.156% for the original one versus 0.234% for the newer one).

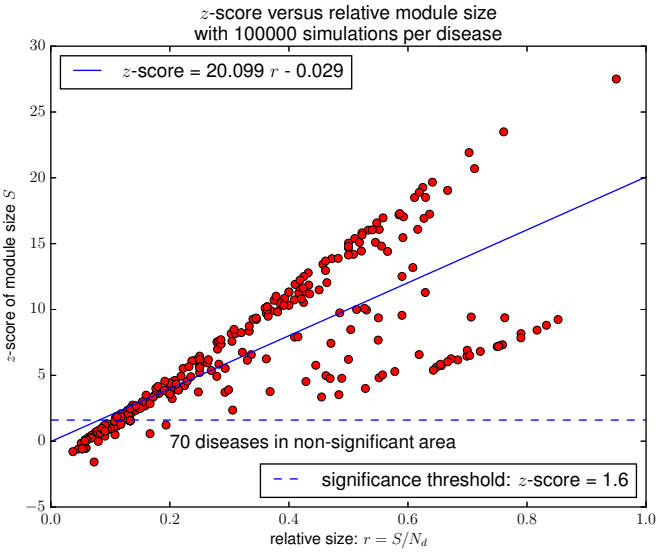

Figure 3: z-score of the largest connected component size vs relative module size of the newer interactome.

Figure 3 is the adaptation of Figure 1 for the newer interactome. We observe that 12 more diseases have a z-score below the threshold of 1.6, which makes results less significant, whereas general relative size has shifted towards right as seen in Figure 4.

This is explained by the increase in the gene number, leading to a higher coverage of the disease genes. Also, z-score maximum has dropped from 31.6 to 27.5, the mean z-score has increased from 6.2 to 6.4 due in part to the density increase in the region z-score $\in [10, 20]$.

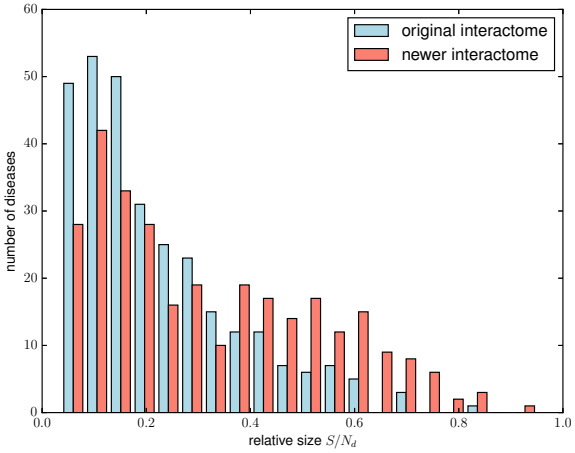

Figure 4: Comparison of relative size distribution between original and newer interactomes.

As well, average relative size for the given diseases has increased from 22% to 32% in the newer interactome, showing the better coverage allowed by the more recent version.

The interpretation of Subsection 2.1 still stand: diseases having a low z-score still have a low relative size, and diseases with a higher z-score have a higher relative size. So either the interactome is still too incomplete for these diseases, or they lie in very sparse regions of the interactome.

Due to the density increase, the degree distribution has changed as well. Figure 5 shows the degree distribution of both the original interactome and the new one. The most visible change is that the newer interactome contains more connected nodes which implies a mean twice as big.

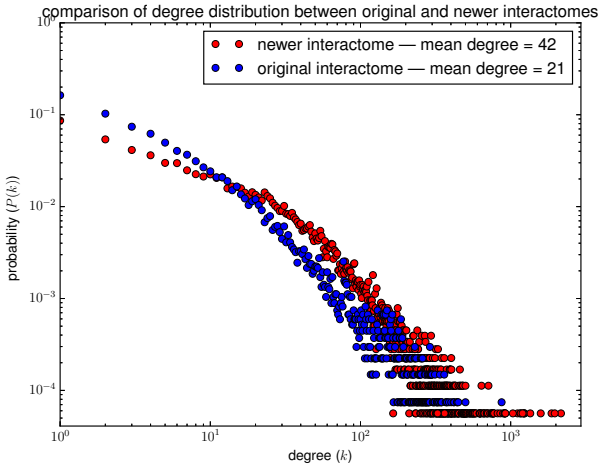

Figure 5: Degree distribution comparison.

Yet, the network is still scale-free, and its degree

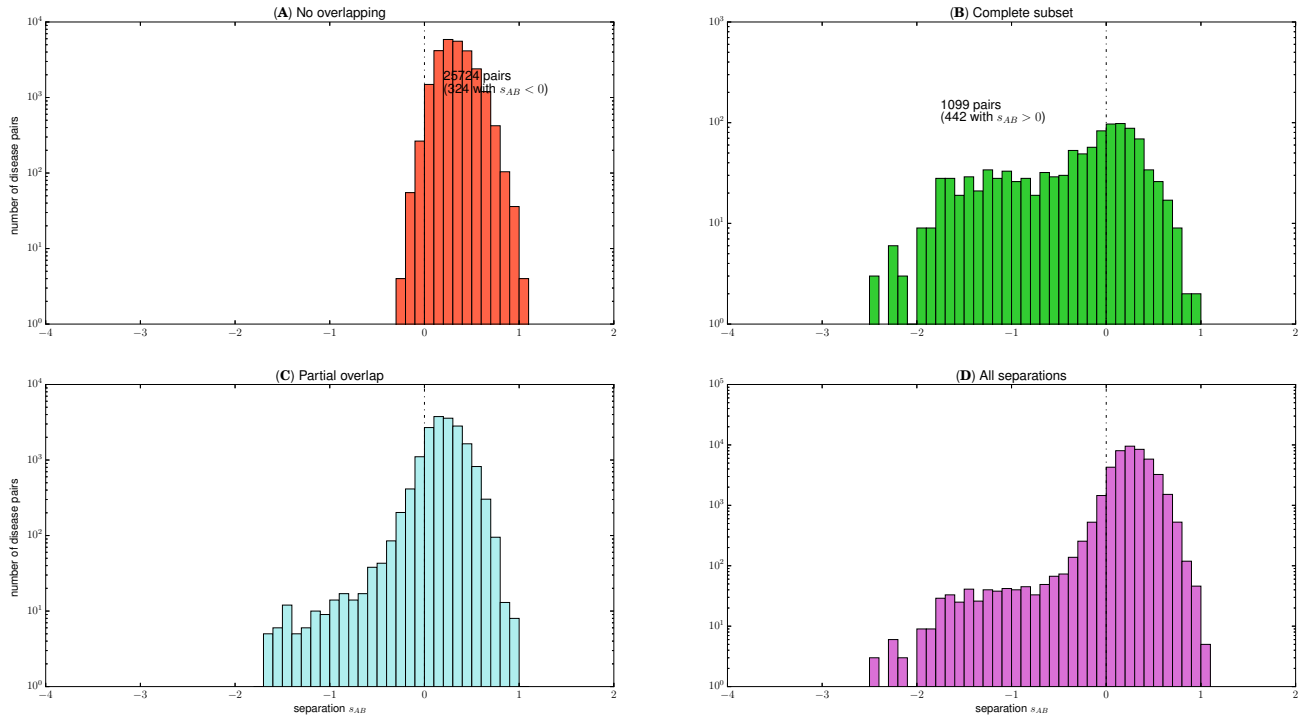

Figure 6: Disease pairs separation in the new interactome.

distribution still follows a power law, which is inherent to biological networks as explained in Vidal et al. (2011) (coefficient 1.6 for the newer one versus 1.53 for the original one) as described in Barabasi and Oltvai (2004) and Seebacher and Gavin (2011).[2] γ is bigger than 1 and smaller than 3, which is considered relevant in Barabasi and Oltvai (2004).

The 12 diseases which have a decrease of z-score below 1.6 is due to the increase of the interactome density implying that a subgraph taken at random tends to have a wider LCC at equal size.

Separation analysis applied on the newer interactome yields results presented in Figure 6. We observe therefore that non-overlapping diseases have a higher separation score in the newer interactome: from 710 disease pairs to 324 pairs having a $s_{AB} > 0$ score (so 54% of the non-overlapping disease pairs increased their separation score above 0).

We also observe that 6% of the complete subset disease pairs decreased their separation score below 0, which is way less significant.

More generally, separation scores have tightened around 0: s_{AB} is in $[-3.2, 1.6]$ in the original interactome, and in the newer one, s_{AB} is in $[-2.5, 1.1]$.

[2]Determined by taking coefficient γ in the relation $\log(P(k)) \sim -\gamma \log(k)$ found by a linear regression.

4 Improvements

4.1 Subgraph largest connected component distribution

Figures 1 and 3 require a null hypothesis in order to define a z-score, being the random one. Those are computed as follows: if S_D is the disease module associated with a given disease D, then its z-score is given by:

$$z\text{-score} = \frac{|S_D| - \mu(S^{\text{rand}})}{\sigma(S^{\text{rand}})}, \tag{1}$$

with $\mu(S^{\text{rand}})$ and $\sigma(S^{\text{rand}})$ being respectively the mean and the standard deviation of the largest connected component size of a random subgraph of size $|D|$ in the interactome.

These values are obtained by simulations: taking subgraphs at random of given size in the interactome. With 10^3 simulations per subgraph size, Figure 7 plots simulated mean and standard deviations of largest connected component size versus subgraph size.

4.2 Analytically determined probability density

In order to avoid simulation computation time, a probability mass function has been determined. For a graph $\Gamma = (V, E)$ such that $|E| = m$ and $\Lambda_k^m(V, \cdot)$, the set of all graphs having V as vertex set, m edges, and a LCC of size k, we define p_k,

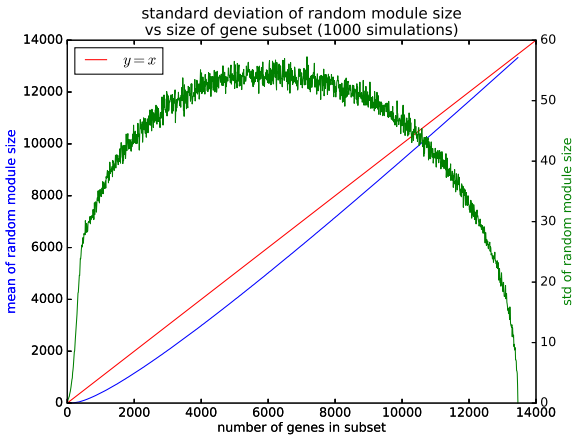

Figure 7: S^{rand} mean and standard deviation distribution.

the probability that Γ has LCC of size k as:

$$p_k = \left|\Lambda_k^m(V,\cdot)\right| \Big/ \binom{\binom{|V|}{2}}{m}, \tag{2}$$

which requires $\left|\Lambda_k^m(V,\cdot)\right|$ to be computed. Yet, this set cardinality is defined by a recurrence relation (see proof in supplementary materials), which makes computations several orders of magnitude slower (even with dynamic programming and caching).

A Python3 implementation is given in `source/lcc_size/`.

5 Conclusion

References

Amaral, L. A. N. (2008). A truer measure of our ignorance. *Proceedings of the National Academy of Sciences*, 105(19):6795–6796.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101.

Barraez, D., Boucheron, S., and Fernandez De LaVega, W. (2000). On the fluctuations of the giant component. *Comb. Probab. Comput.*, 9(4):287–304.

Catabia, H., Smith, C., and Ordovás, J. (2017). Inter-tools: a toolkit for interactome research.

Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2015). Dida: A curated and annotated digenic diseases database. *Nucleic acids research*, 44(D1):D900–D907.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., et al. (2007). Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. *BMC biology*, 5(1):44.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224).

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., Rechenmann, F., and Jacq, B. (1999). Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. *Nucleic acids research*, 27(1):89–94.

Seebacher, J. and Gavin, A.-C. (2011). Snapshot: Protein-protein interaction networks. *Cell*, 144(6):1000–1000.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964.

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.

Yu, L., Wang, B., Ma, X., and Gao, L. (2016). The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC systems biology*, 10(4):111.