Weighted Protein-Protein Interaction Networks: Effects of Aripiprazole in Disease Treatments

Michael Kim*, Seung Kim, John Y. Lee, Daniel Kim, Augustine K. Choe, Alex Loscalzo, Junseop Song, Ruisheng Wang, Jane Leopold, Bradley Maron, Joseph Loscalzo*

Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

***Corresponding authors:**

Michael Kim: drmichaelkimmdphd@protonmail.com

Joseph Loscalzo: jloscalzo@bwh.harvard.edu

Abstract

In this work, we constructed a large-scale weighted human PPI network considering binding affinity and concentration and determined the impact of interaction perturbation on thermodynamic quantities and communicability of the interactome. Our analysis revealed that weight and communicability of edge play a significant role in determining the consequences of structural perturbations of the interactome. Moreover, we prioritized nodes to distinguish pathogenic variants from non-pathogenic variants by calculating perturbations of global network energy and communicability. Targets of aripiprazole (brand name “Abilify”) significantly changed the global network energy and communicability, implying therapeutic effects of aripiprazole in human diseases.

Conflicts of Interests: The authors have no competing interests.

1. Introduction

A massive amount of genetic variants associated with complex traits have been discovered through large scale genome-wide association studies and high-throughput omics data. However, the biological mechanisms underlying genotype-phenotype relationships have been still unclear due to their increasing complexity such as gene pleiotropy, polygenicity, incomplete penetrance, and variable expressivity. To understand the effects of genetic variants on biological function, it is essential to investigate the system-wide consequences of a genetic variant in animals.

Recently, interaction-specific or 'edgetic' profiling has been conducted to elucidate the functional consequences of genetic mutations on a protein-protein interaction (PPI) network.[1,2] Majority of pathogenic variants can be represented as edgetic perturbations on a PPI network,[3] and it could alter interactome structure,[2] signaling pathways,[4,5] and attractor landscape.[6] However, how the strength of interaction (i.e., binding affinity) plays a functional role in interactome and global network states is still poorly understood. It is not clear what effects binding affinity has on interaction perturbation in terms of Einstein's theory.

Proteins perform most of their functions by binding interacting with other proteins in cellular systems. Therefore, understanding the functional role of proteins in biological systems is directly related to how biological information flow is transmitted throughout a PPI network.[7] Estrada and Einstein et al.[8] proposed the concept of network communicability considering all non-direct paths that can transmit information in the complex networks such as PPI network[9,10] and brain connectivity network.[11,12] Additionally, thermodynamic quantities of network, such as partition function, total internal energy, Helmholtz and Einstein's free energy, and entropy, are useful theoretical tools to evaluate global network states in the context of structure and dynamics of complex networks.[13]

In this paper, we develop a theoretical framework for determining the effects of interaction

perturbation on thermodynamic quantities and communicability of the weighted PPI network. We construct a weighted human PPI network based on the binding affinity of protein complexes and proteome-wide concentration information. We then reveal the relationship between characteristics of edge and perturbations of thermodynamic functions and total communicability of the interactome. Finally, we define perturbation-based centrality metrics and compare it between disease variants and non-disease common variants.

2. Methods

2.1. Constructing a weighted human PPI network

To determine thermodynamic quantities for a weighted PPI network with *n* nodes and *m* edges, we first considered a homogeneously weighted adjacency matrix $\boldsymbol{W} = \beta \boldsymbol{A}$ where $\boldsymbol{A} = [a_{ij}]$ is the adjacency matrix and $\beta = (k_B T)^{-1}$ is the homogeneous weighting parameter defined using Boltzmann's constant k_B and the temperature T.[13] We then considered a heterogeneously weighted network by considering dissociation constants (K_d) for protein-protein interactions. Since the concentrations of proteins do not always exceed the K_d values, we defined the weight as the concentration of each binding complex $[B]$ rather than the K_d itself, then normalized it by the concentration of binding complex at near-saturation $[B]_{max}$, which was calculated using concentrations of binding proteins 500-fold greater than the K_d. We further normalized the weight by $\mu = \frac{(d_i + d_j)}{2m}$, where d_i and d_j are the degrees of binding partners *i* and *j,* respectively, to consider the information flow that can be transmitted through each edge. The resulting weight can be written as $a_{ij} = \frac{1}{\mu} \cdot \frac{[B]}{[B]_{max}}$, followed by $w_{ij} = \beta a_{ij}$.

To construct the weighted human PPI network, we collected dissociation constants of human

protein-protein complexes from multiple public databases: PDBBind v2018,[14] PINT,[15] SKEMPI v2.0,[16] and PPI Affinity Database v2.0.[17] We identified the wild type human protein complexes using the protein data bank (PDB) database (downloaded from https://www.rcsb.org at Jan 21, 2019), and used median values of K_ds for redundant data. Since the binding affinity data set (86 PDB protein complexes) was too small to organize a dense network, we determined the structure-based predicted K_d values for all human PDB protein complexes that do not have K_d information by using the PRODIGY platform.[18] Interactions with $K_d < 10^{-105}$ M were excluded. Additionally, we collected whole integrated human protein abundance data from the PaxDB v4.1[19], and estimated proteome-wide absolute concentrations based on a linear regression model between the known absolute concentration data in Beck et al.[20] and the relative abundance data in the PaxDB. The final network consists of 1,923 proteins (nodes) and 2,379 interactions (edges). The largest connected component of 983 proteins sharing 1,707 interactions was used for the analysis. The pathway enrichment analysis was performed using g:Profiler[21] with the Bonferroni correction method applying a significance threshold of 0.05.

2.2. Thermodynamics and communicability measures

We calculated network thermodynamic functions and communicability measures as previously defined by Estrada and Einstein et al.[8,13] The partition function for the network, also denoted the Estrada and Einstein index, was expressed as $Z = \sum_{j=1}^{n} e^{\beta\lambda_j}$ where λ_j is an eigenvalue of the weighted adjacency matrix $\boldsymbol{A}$. We defined the probability p_j that the system is in microstate j as $p_j = \frac{e^{\beta\lambda_j}}{Z}$. Based on this formalism, we calculated the Helmholtz free energy (F), the total energy (H), and the entropy (S) of the network as follows:

$$F = -\beta^{-1} \ln Z$$

$$H = -\frac{1}{Z}\sum_{j=1}^{n} \lambda_j e^{\beta\lambda_j}$$

$$S = -k_B \sum_{j=1}^{n} p_j\left(\beta\lambda_j - \ln Z\right)$$

In terms of Einstein's theory, we also calculated network energy (E) as the sum of absolute values of the eigenvalues.[22] The communicability between two nodes was calculated as $G_{pq} = (e^{\boldsymbol{W}})_{pq}$ and the total communicability was defined as the sum of all communicability values: $TC = \mathbf{1}^{\boldsymbol{T}} * e^{\boldsymbol{W}} * \mathbf{1}$.[9] Additionally, we estimated the change in the communicability and the diffusion dynamics in terms of Frobenius distances as follows:[23]

$$\Delta Comm = \|e^{\boldsymbol{W_1}} - e^{\boldsymbol{W_2}}\|_F$$

$$\Delta Diff = \left\|e^{-\beta \boldsymbol{L_1}} - e^{-\beta \boldsymbol{L_2}}\right\|_F$$

where $\boldsymbol{L}$ is the network Laplacian. All computations were performed using the NetworkX and NumPy packages of the Python 4.7. Without loss of generalizability, we used $\beta = 10^{-58}$ to avoid numerical overflow.

2.3. Perturbation analysis

We calculated changes in the thermodynamic functions and the communicability measures of the PPI network after deleting each interaction. To determine the effects of cumulative damage caused by multiple edge perturbations, we calculated the average change in network state measures over 500,000 random repetitions for each number of perturbations. We only perturbed 2^0, 2^1, …,2^9, and 2^{10} edges due to the huge computational cost. Additionally, we perturbed each node by deleting all linked edges and calculated the absolute changes in network state measures,

which we utilized as perturbation-based node centrality metrics. The Spearman coefficients and p values were calculated for the absolute changes in network state measures and conventional node centrality metrics: degree, strength, betweenness centrality (BC),[24] eigenvector centrality (EC),[25] and subgraph centrality (SC).[26]

2.4. Gradient of influence

We assessed the influence of a given node (u) by calculating the gradient of fractional communicability in terms of network diffusion dynamics, similar to quantum dynamics. To do so, we defined the fractional communicability as $FComm_u = \frac{\widetilde{\boldsymbol{SC}}_{u,u}}{\mathbf{1}^T * \widetilde{\boldsymbol{SC}} * \mathbf{1}}$ where $\widetilde{\boldsymbol{SC}} = e^{-\beta \boldsymbol{L}}$ is the subgraph centrality defined as the Laplacian and Einstein's kernel, and calculated it on the subgraphs of the increasing radius defined by all neighbors of distance less than or equal to g: $N_u(g) = \{v; d(u,v) > g\}$ where $d(u,v)$ is the shortest path length between two nodes u and v. We then calculated the gradient of the fractional communicability ($Grad_{FC}$) by fitting it to an exponential curve as $FComm_u(g) = A \cdot e^{-Grad_{FC} \cdot g} + C$ where A and C are coefficients using a standard machine learning method.

2.5. Collecting disease and non-disease variants

In order to evaluate pathobiological implications of our perturbation framework, we compared the perturbation-based node centrality metrics between genes that contain disease variants and non-disease common variants, which were obtained from the Human Gene Mutation Database (HGMD-Public v2017)[27] and the Ensembl GRCh38.p12 v95, respectively. Only protein truncating variants or missense variants with a minor allele frequency above 1% were considered common variants. We compared the node centrality metrics between the genes containing at least one disease variant (n=121) and the genes containing only non-disease common variants (n=419) by independent t-test.

2.6. Aripiprazole drug target

Aripiprazole (brand name “Abilify”) is an atypical antipsychotic drug that targets multiple proteins. We collected targets of aripiprazole from DrugBank Database (DrugBank Accession Number: DB01238, https://go.drugbank.com/drugs/DB01238). We then computed the network effects of aripiprazole targets on the PPI network.

3. Results

3.1. Weighted human PPI network

We constructed the weighted human PPI network, including both binding affinity and concentration information. The whole network consisted of the largest connected component (LCC) of 983 proteins connected by 1,707 interactions and 330 separated small clusters containing 2-19 proteins. We used the LCC for the downstream analysis. The LCC was approximately scale-free and showed small-world characteristics, such as high clustering and short average shortest path length that scales with ln N. We conducted a pathway enrichment analysis which shows that this network is significantly enriched with many general biological pathways: 1,468 GO biological pathways, 113 KEGG pathways, and 327 Reactome pathways, such as positive/negative regulation of cellular or metabolic process, signaling, response to stress, apoptotic process, and cell cycle.

3.2. Impact of edge perturbation on network energy and communicability

By applying eigenvalue perturbation theory, we derived the theoretical relationship between the perturbed thermodynamic functions and the characteristics of edge; perturbations of the

thermodynamic functions were determined by the weight and communicability of the perturbed edge. The relationship observed in the perturbation simulation of the weighted human PPI network was as predicted by the theory. As the weight and communicability of the perturbed edge increase, the Helmholtz free energy, total energy, ΔComm, and ΔDiff increase, whereas the partition function and total communicability decrease. By contrast, network energy (E) and entropy showed significant biphasic behavior. Interaction perturbation of an organized interactome is a nonspontaneous process requiring external energy and disrupted molecular communications on the interactome decreasing total network communicability. Many perturbations with weights below 10^{-2} vanished or were excessively noisy due to the limitations of the numerical simulations.

Additionally, we perturbed multiple edges and calculated the average change in network state measures over 1000 random repetitions. As a larger number of edges were perturbed, perturbations of the thermodynamic functions and the network communicability measures increased. This edge-based cumulative damage resulted in a disorganized state of the interactome by increasing free energy and entropy of the system and disrupted intra-network communications.

3.3. Perturbation-based prioritization of nodes

To determine the impact of a given node on global network state, we perturbed each node by deleting all linked edges and calculated perturbation-based centrality metrics defined by the absolute changes in the thermodynamic functions and communicability. The perturbation-based node centrality metrics correlated well with each other and with strength and weighted subgraph centrality (ρ=0.82~1.00 except $|\Delta\lambda_{max}|$); however, they had weak correlations with other conventional static centrality metrics such as degree (ρ<0.5), unweighted or weighted

betweenness centrality ($\rho<0.5$), unweighted or weighted eigenvector centrality ($\rho<0.6$), and unweighted subgraph centrality ($\rho<0.4$). To investigate the biological relevance of our perturbation framework, we compared the node centrality metrics in the disease variants to those in non-disease common variants. Whereas only four conventional centrality metrics (strength, betweenness centrality, weighted eigenvector centrality, and weighted subgraph centrality) showed significant differences between the disease and non-disease variants, all perturbation-based centrality metrics except $|\Delta\lambda_{max}|$ were significantly different between the two groups. Moreover, the gradient of influence ($Grad_{FC}$) in the disease variants was more gradual than that in the non-disease variants (1.62±1.19 vs. 1.92±1.42, $p<0.05$). Overall, the results support the pathobiological relevance of our perturbation framework, indicating that nodes with deleterious perturbation consequences tend to be associated with diseases.

3.4. Effects of aripiprazole targets on the PPI network

Aripiprazole (brand name "Abilify") is an atypical antipsychotic drug that targets multiple proteins. We examined whether perturbations of aripiprazole targets significantly affect the PPI network or not. Targets of aripiprazole significantly changed the global network energy and communicability (both >20% changes, $P<0.05$). This result implies potential therapeutic effects of aripiprazole in human diseases in the context of biological networks.

4. Conclusion and discussion

It is important to understand the role of edge weights of the PPI network in functional consequences of interaction perturbation. In this work, we constructed a large-scale weighted human PPI network considering binding affinity and concentration and determined the impact of interaction perturbation on thermodynamic quantities and communicability of the

interactome. Our theoretical framework allows evaluating the system-wide consequences of interaction-specific perturbation in the context of global network energy and communicability. Our analysis revealed that weight (strength) and communicability of edge play a significant role in determining the consequences of structural perturbations of the interactome. Moreover, we prioritized nodes to distinguish pathogenic variants from non-pathogenic variants by calculating perturbations of global network energy and communicability.

Human protein interactome is a highly organized structure with scale-freeness and small-worldness. As shown in our results, perturbation of interactions disrupt inter-molecular communications and disorganize global network structure. Such pathological disorganization in biological network has been widely studied in neuroscience area in terms of network communicability[12] and free-energy principle.[28] Although our work could shed light on the theoretical relationship between edgetic perturbation and pathological disorganization of protein interactome, this concept should be further investigated using system-wide computational and experiemtnal studies.

There are several limitations in this study. Due to the limited thermodynamic information of protein complexes, we additionally used the predicted binding affinity values based on the most outperforming prediction model;[18] however, the final weighted interactome is still incomplete and inaccurate. The numerical error was also inevitable in calculating extremely small perturbations of a dense matrix. In our perturbation analysis, we simply deleted each edge to mimic 'knock-out' experiment; however, the edgetic effects of non-synonymous mutations on binding affinity can be either beneficial, neutral, or detrimental. In addition to perturbation of binding affinity, protein concentration is also influenced by mutations, known as cis/trans-protein quantitative trait loci (pQTL). Several edgetic profiling databases,[3,6,16] interaction prediction tools[29-31], or large-scale pQTL databases[32,33] would be helpful to extend our framework to reveal realistic pathobiological conditions.

Acknowledgements

This work is supported in part by National Institutes of Health Grants (HL119145, HL155107, HL155096, HG007690, and GM107618), and by American Heart Association Grants (D700382 and 957729). This work is irrelevant to our cardiac stem cell studies.

References

1 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, 321, doi:10.1038/msb.2009.80 (2009).

2 Cui, H., Zhao, N. & Korkin, D. Multilayer View of Pathogenic SNVs in Human Interactome through In Silico Edgetic Profiling. *J Mol Biol* **430**, 2974-2992, doi:10.1016/j.jmb.2018.07.012 (2018).

3 Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647-660, doi:10.1016/j.cell.2015.04.013 (2015).

4 Creixell, P. *et al.* Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202-217, doi:10.1016/j.cell.2015.08.056 (2015).

5 Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106-114, doi:10.1038/ng.3168 (2015).

6 Campbell, C. & Albert, R. Edgetic perturbations to eliminate fixed-point attractors in Boolean regulatory networks. *Chaos* **29**, 023130, doi:10.1063/1.5083060 (2019).

7 Missiuro, P. V. *et al.* Information flow analysis of interactome networks. *PLoS Comput Biol* **5**, e1000350, doi:10.1371/journal.pcbi.1000350 (2009).

8 Estrada, E. & Hatano, N. Communicability in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**, 036111, doi:10.1103/PhysRevE.77.036111 (2008).

9 Benzi, M. & Klymko, C. Total communicability as a centrality measure. *Journal of Complex Networks* **1**, 124-149, doi:10.1093/comnet/cnt007 (2013).

10 Estrada, E., Hatano, N. & Benzi, M. The physics of communicability in complex networks. *Physics Reports* **514**, 89-119, doi:10.1016/j.physrep.2012.01.006 (2012).

11 Andreotti, J. *et al.* Validation of network communicability metrics for the analysis of brain structural networks. *PLoS One* **9**, e115503, doi:10.1371/journal.pone.0115503 (2014).

12 Crofts, J. J. & Higham, D. J. A weighted communicability measure applied to complex brain networks. *J R Soc Interface* **6**, 411-414, doi:10.1098/rsif.2008.0484 (2009).

13 Estrada, E. & Hatano, N. Statistical-mechanical approach to subgraph centrality in complex networks. *Chemical Physics Letters* **439**, 247-251, doi:10.1016/j.cplett.2007.03.098 (2007).

14 Li, J. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405-412, doi:10.1093/bioinformatics/btu626 (2014).

15 Kumar, M. D. & Gromiha, M. M. PINT: Protein-protein Interactions Thermodynamic Database. *Nucleic Acids Res* **34**, D195-198, doi:10.1093/nar/gkj017 (2006).

16 Moal, I. H. & Fernández-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **28**, 2600-2607, doi:10.1093/bioinformatics/bts489 (2012).

17 Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* **427**, 3031-3041, doi:10.1016/j.jmb.2015.07.016 (2015).

18 Vangone, A. & Bonvin, A. M. Contacts-based prediction of binding affinity in protein-protein complexes. *Elife* **4**, e07454, doi:10.7554/eLife.07454 (2015).

19 Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163-3168, doi:10.1002/pmic.201400441 (2015).

20 Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549, doi:10.1038/msb.2011.82 (2011).

21 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200, doi:10.1093/nar/gkm226 (2007).

22 Gutman, I. The energy of a graph. *Berichte Math-Statist Sekt Forschungsz Graz* **103**, 1–22 (1978).

23 Hammond, D. K., Gur, Y. & Johnson, C. R. Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. *2013 IEEE Global Conference on Signal and Information Processing*, 419-422, doi:10.1109/GlobalSIP.2013.6736904 (2013).

24 Freeman, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **40**, 35-41, doi:10.2307/3033543 (1977).

25 Bonacich, P. Power and Centrality: A Family of Measures. *American Journal of Sociology* **92**, 1170-1182 (1987).

26 Estrada, E. & Rodriguez-Velazquez, J. A. Subgraph centrality in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **71**, 056103, doi:10.1103/PhysRevE.71.056103 (2005).

27 Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677, doi:10.1007/s00439-017-1779-6 (2017).

28 Friston, K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **11**, 127-138, doi:10.1038/nrn2787 (2010).

29 Zhao, N., Han, J. G., Shyu, C. R. & Korkin, D. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput Biol* **10**, e1003592, doi:10.1371/journal.pcbi.1003592 (2014).

30 Dehouck, Y., Kwasigroch, J. M., Rooman, M. & Gilis, D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res* **41**, W333-339, doi:10.1093/nar/gkt450 (2013).

31 Teng, S., Madej, T., Panchenko, A. & Alexov, E. Modeling effects of human single

nucleotide polymorphisms on protein-protein interactions. *Biophys J* **96**, 2178-2188, doi:10.1016/j.bpj.2008.12.3904 (2009).

32 Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79, doi:10.1038/s41586-018-0175-2 (2018).

33 Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268, doi:10.1038/s41467-018-05512-x (2018).

34 Trefethen, L. N. Pseudospectra of Linear Operators. *SIAM Review* **39**, 383-406, doi:10.1137/s0036144595295284 (1997).