G2G: A web-server for the prediction of human synthetic lethal interactions

Yom Tov Almozlino [a], Iftah Peretz [b], Martin Kupiec [b], Roded Sharan [a,*]

[a] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
[b] School of Molecular Cell Biology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

ABSTRACT

Genetic interactions (GIs) are fundamental to our understanding of biological processes in the cell. While GIs have been systematically mapped in yeast, there is scarce information about them in humans. Recently, we have suggested a state-of-the-art hierarchical method that leverages gene ontology information for predicting GIs in yeast. Here, we adapt this method and apply it for the first time to predict GIs in human. We introduce a web service called G2G for this task that is available at http://bnet.cs.tau.ac.il/g2g/.

© 2020 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

As genes never work in isolation, their effects are oftentimes affected by the activity of other genes. Thus, the phenotype of a certain mutation can be affected by mutations in other genes, such that the cell's fitness may be further reduced (negative, synthetic lethal (SL) interactions) or improved (positive interactions) [1]. Dissecting the genetic interactions of particular genes provides information about the processes in which the genes participate, and provides a high-order linkage between cellular processes. A systematic search for positive and negative genetic interactions has been carried out in many model organisms, such as E. coli [2] and yeast [3] providing important information about the wiring within the prokaryotic and eukaryotic cells.

Genetic interactions (GIs) play an important role in the etiology of many human diseases [4–8], thus having potential applications in their diagnosis and treatment. However, systematic genetic interaction maps are hard to construct and relatively little is known about the human genetic interactome.

Previous work focused mainly on yeast, where systematic GI data is available. It includes the use of flux balance analysis to simulate the impact of gene deletions on cell growth [9]; Guilt-by-association, which predicts the phenotype of pairwise gene deletions based on the phenotypes of their network neighbors [10]; the Multi-network multi-classifier, a 'black box' supervised learning system which uses many different lines of experimental evidence as features to predict genetic interactions [11]; and

recently the Ontotype method (sketched below) which leverages Gene Ontology information to predict genetic interactions [12].

In humans, RNA interference and CRISPR screens were used to detect GIs either systematically or by targeting specific driver mutations [13–17]. Computationally, various methods were developed for GI prediction, including models that prioritize human GIs based on experimental information on their yeast orthologs [18,19], the MiSL method for predicting SL partners of specific driver mutations by searching for genes that are either amplified or not deleted in the presence of those mutations [20], the discover SL method which predicts SL interactions from mutation, gene expression and copy number data [21], similarly a model that predicts SL interactions based on the absence of co-loss of these genes in expression or copy-number data [22], and the ISLE method [23], which screens for gene pairs that: (i) display significantly less frequent co-inactivation than expected in tumor sample, (ii) upon co-inactivation improve patients' survival, and (iii) have similar phylogenetic profiles as curated from 86 species. Additional prediction methods are based on matrix factorization [24], network topology [25] and integration of heterogeneous data sources [26].

Here we present the G2G web service (http://bnet.cs.tau.ac.il/g2g/) that allows users to predict phenotypes of pairwise gene deletions in *Homo sapiens* using a streamlined implementation of the algorithm described by Yu et al. [12] and a novel extension of it. The algorithm is based on the concept of an *ontotype*, an intermediate representation between genotype and phenotype which is defined by the gene ontology terms associated with the genes that are mutated in the genotype. By computing the ontotype of a large set of experimentally measured genetic interactions, one may construct a supervised learning model using the term columns as

features and the measurements as target values. Such a model can predict the expected phenotype of any given pair of genes.

2. Methods

The original algorithm in [12] was designed for *Saccharomyces cerevisiae*, where systematic knockout data for over 23 million gene pairs is available. G2G adapts the algorithm for *Homo sapiens*, following the same general outline but using random forest classification instead of regression. To train its prediction model, we use the dataset of [17] which has measurements on the deletion phenotype of 105,360 gene pairs in the K562 cell line, including 1678 that are classified as "negative" (synthetic lethal, average GI score ≤ -3) and 690 that are classified as "positive" (average GI score ≥ 3). We construct a training set comprised of the negative gene pairs and a randomly selected equally-sized set of "neutral" (average score between -3 and 3) gene pairs. We use the resulting ontotypes to train a predictor which can estimate the probability that a given gene pair is synthetic lethal.

As the training data in *Homo sapiens* is of small magnitude compared to yeast, we thought to extend the method by enriching the ontotype vectors with information from proteins that are significantly close to the deleted genes in the protein–protein interaction network. To this end, we added a preprocessing step to the prediction pipeline in which we first transform every genotype vector using network propagation. In this process, the two deleted genes are viewed as sources of "heat", which is iteratively diffused through the network. As a result, every protein in the network ends up with a score that reflects its proximity to the initial genes. We use the approach in [27] to assess the significance of these scores, and add the proteins which pass an FDR threshold to the genotype of the initial pair. The computation is performed using the protein–protein interaction network described in [28] which is based on the BioGRID [29] and IntAct [30] interaction databases. G2G can be configured to incorporate the propagation as an additional step before ontotype generation. The implementation uses a default propagation alpha of 0.8 and FDR threshold of 0.05.

G2G is implemented as a Flask server in Python using the standard Pandas and Numpy libraries [31,32] to read the training data and construct the ontotypes. We used the GOATOOLS library [33] to work with the gene ontology and gene annotation files. The supervised learning models are generated using scikit-learn [34].

3. Usage

G2G offers two main modes of operation (see Fig. 1). Users can submit both a source gene and a target gene to compute the phenotype for that pair. Alternatively, they can submit just a source gene, in which case G2G returns all predicted genetic interactions between the gene and its direct neighbors in the protein–protein

Fig. 1. Input parameters for G2G. Source gene, target gene (optional), propagation toggle and a score threshold.
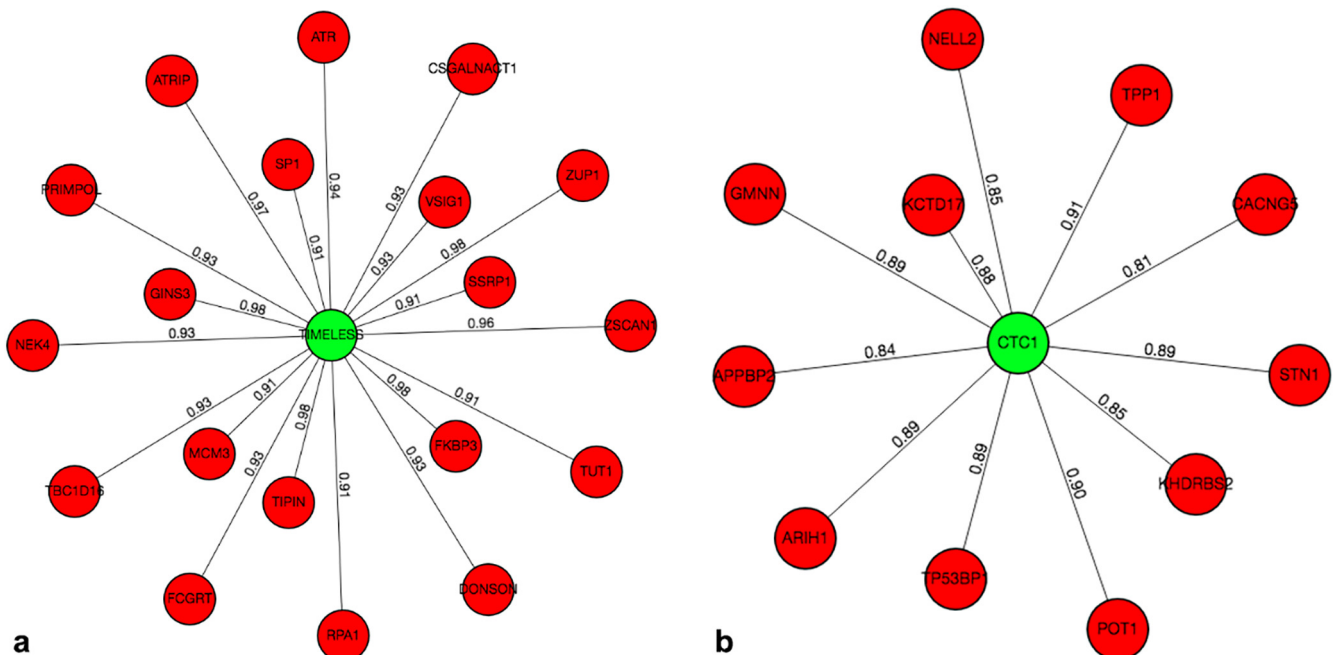

Fig. 2. G2G graphical outputs for source genes TIMELESS (a) and CTC1 (b). The center node, in green, represents the source gene and each of its predicted neighbors, in red, of the PPI network. Each edge of the graph is labeled with the phenotype predicted score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interaction network. In this mode, users can specify a threshold between 0.0 and 1.0 in order to filter out pairs whose phenotype predictions are lower than the threshold. Genes can be entered using their entrez gene ids, official symbols or synonyms as defined by NCBI.

G2G returns its results both as a table listing each gene pair with its phenotype score and as a graph visualization with the genes as nodes (source in green, targets in red) and the phenotype scores as edge labels (see Fig. 2).

4. Case study

To illustrate the use of the web-server, we describe two applications of it to elucidate the genetic interactions of genes for which some literature knowledge is available. As a first example, we selected the TIMELESS (Tim) gene. TIMELESS is known for its role in the circadian clock mechanism in *Drosophila* [35–37], whereas in yeast and humans the TIMELESS-TIPIN protein complex has been reported to be important for replication checkpoint and normal DNA replication processes [38].

When executing a G2G job with TIMELESS as its source gene, a threshold of 0.9 and the propagate checkbox indicated, we get the TIPIN as the target gene with a 0.976 predicted Phenotype score. Furthermore, the system predicts interactions with a number of DNA replication checkpoint genes (e.g.: ATR, ATRIP), as well as a number of DNA replication machinery components, such as RPA1, PRIMPOL, GINS3 and MCM3. These results are supported by the reported interaction of TIMELESS with MCM2-7 during DNA replication [39,40].

As a second example, a search for the telomeric protein CTC1 with a threshold of 0.6 returned other telomeric proteins, such as STN1, POT1, TPP1 [41] as well as additional proteins that respond to telomere DNA damage, such as 53BP1 [42] (Fig. 2).

5. Performance evaluation

In addition to the case studies, we systematically evaluated G2G's performance by a stratified 6-fold cross-validation test using the dataset of [17]. We report the area under the ROC curve (AUC) as this is a commonly used measure for such tasks that is robust to label imbalance [43].

Over all folds, G2G achieved a mean AUC of 0.76 with a small standard deviation of 0.04. The results slightly improved when using the propagation feature, yielding a mean AUC of 0.77 (±0.03).

6. Conclusions

We have presented the G2G web-server for predicting and visualizing human synthetic lethal interactions. We have shown the agreement of the predictions with the literature across two independent case studies, as well as the good performance in cross validation on systematic phenotypic data. G2G can be used to score specific putative interactions or to prioritize potential GIs that overlap the protein–protein interactions of a gene of interest. We expect G2G to assist researchers in the characterization of GIs and their clinical applications.

CRediT authorship contribution statement

Yom Tov Almozlino: Methodology, Software, Visualization, Writing - original draft. **Iftah Peretz:** Validation, Writing - review & editing. **Martin Kupiec:** Validation, Writing - review & editing. **Roded Sharan:** Conceptualization, Supervision, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

[1] Gilbert-Diamond D, Moore JH. Analysis of Gene-Gene Interactions. Curr Protocols Hum Genet 2011. https://doi.org/10.1002/0471142905.hg0114s70.

[2] Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, et al. Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia coli. PLoS Genet 2014;10: e1004120.

[3] Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science 2016;353. https://doi.org/10.1126/science.aaf1420.

[4] Yang Q, Khoury MJ, Sun F, Flanders WD. Case-only design to measure gene-gene interaction. Epidemiology 1999;10:167–70.

[5] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009;10:392–404.

[6] Ma DQ, Whitehead PL, Menold MM, Martin ER, Ashley-Koch AE, Mei H, et al. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. Am J Hum Genet 2005;77:377–88.

[7] Meng Y, Groth S, Quinn JR, Bisognano J, Wu TT. An Exploration of Gene-Gene Interactions and Their Effects on Hypertension. Int J Genomics 2017;2017:1–9. https://doi.org/10.1155/2017/7208318.

[8] Hoppe C, Klitz W, Cheng S, Apple R, Steiner L, Robles L, et al. Gene interactions and stroke risk in children with sickle cell anemia. Blood 2004;103:2391–6.

[9] Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 2011;43:656–62.

[10] Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 2011;21:1109–21.

[11] Pandey G, Zhang B, Chang AN, Myers CL, Zhu J, Kumar V, et al. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Comput Biol 2010;6. https://doi.org/10.1371/journal.pcbi.1000928.

[12] Yu MK, Kramer M, Dutkowski J, Srivas R, Licon K, Kreisberg J, et al. Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. Cell Syst 2016;2:77–88.

[13] Vizeacoumar FJ, Arnold R, Vizeacoumar FS, Chandrashekhar M, Buzina A, Young JTF, et al. A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. Mol Syst Biol 2013;9:696.

[14] Scholl C, Fröhling S, Dunn IF, Schinzel AC, Barbie DA, Kim SY, et al. Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell 2009;137:821–34.

[15] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108–12.

[16] Shen JP, Zhao D, Sasik R, Luebeck J, Birmingham A, Bojorquez-Gomez A, et al. Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. Nat Methods 2017;14:573–6. https://doi.org/10.1038/nmeth.4225.

[17] Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, et al. Mapping the Genetic Landscape of Human Cells. Cell 2018;174. 953–67.e22.

[18] Deshpande R, Asiedu MK, Klebig M, Sutor S, Kuzmin E, Nelson J, et al. A comparative genomic approach for identifying synthetic lethal interactions in human cancer. Cancer Res 2013;73:6128–36.

[19] Srivas R, Shen JP, Yang CC, Sun SM, Li J, Gross AM, et al. A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy. Mol Cell 2016;63:514–25.

[20] Sinha S, Thomas D, Chan S, Gao Y, Brunen D, Torabi D, et al. Systematic discovery of mutation-specific synthetic lethals by mining pan-cancer human primary tumor data. Nat Commun 2017;8:15580.

[21] Das S, Deng X, Camphausen K, Shankavaram U. DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers. Bioinformatics 2019;35:701–2.

[22] Lu X, Megchelenbrink W, Notebaart RA, Huynen MA. Predicting human genetic interactions from cancer genome evolution. PLoS ONE 2015;10:e0125795.

[23] Lee JS, Das A, Jerby-Arnon L, Arafeh R, Auslander N, Davidson M, et al. Harnessing synthetic lethality to predict the response to cancer treatment. Nat Commun 2018;9:2546.

[24] Liu Y, Wu M, Liu C, Li X, Zheng J. SL 2 MF: Predicting Synthetic Lethality in Human Cancers via Logistic Matrix Factorization. IEEE/ACM Trans Comput Biol Bioinform 2019. https://doi.org/10.1109/TCBB.2019.2909908.

[25] Benstead-Hume G, Chen X, Hopkins SR, Lane KA, Downs JA, Pearl FMG. Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. PLoS Comput Biol 2019;15:. https://doi.org/10.1371/journal.pcbi.1006888e1006888.

[26] Liany H, Jeyasekharan A, Rajan V. Predicting synthetic lethal interactions using heterogeneous data sources. Bioinformatics 2020;36:2209–16.

[27] Biran H, Almozlino T, Kupiec M, Sharan R. WebPropagate: A Web Server for Network Propagation. J Mol Biol 2018;430:2231–6.

[28] Almozlino Y, Atias N, Silverbush D, Sharan R. ANAT 2.0: reconstructing functional protein subnetworks. BMC Bioinf 2017;18:495.

[29] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9.

[30] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63.

[31] McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 2010. Doi: 10.25080/majora-92bf1922-00a.

[32] van der Walt S, van der Walt S, Chris Colbert S, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng 2011;13:22–30. https://doi.org/10.1109/mcse.2011.37.

[33] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 2018;8:10872.

[34] Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

[35] Sehgal A, Price JL, Man B, Young MW. Loss of circadian behavioral rhythms and per RNA oscillations in the Drosophila mutant timeless. Science 1994;263:1603–6.

[36] Vosshall LB, Price JL, Sehgal A, Saez L, Young MW. Block in nuclear localization of period protein by a second clock mutation, timeless. Science 1994;263:1606–9.

[37] Sehgal A, Rothenfluh-Hilfiker A, Hunter-Ensor M, Chen Y, Myers MP, Young MW. Rhythmic expression of timeless: a basis for promoting circadian cycles in period gene autoregulation. Science 1995;270:808–10.

[38] Leman AR, Noguchi C, Lee CY, Noguchi E. Human Timeless and Tipin stabilize replication forks and facilitate sister-chromatid cohesion. J Cell Sci 2010;123:660–70.

[39] Xu X, Wang J-T, Li M, Liu Y. TIMELESS Suppresses the Accumulation of Aberrant CDC45·MCM2-7·GINS Replicative Helicase Complexes on Human Chromatin. J Biol Chem 2016;291:22544–58.

[40] Numata Y, Ishihara S, Hasegawa N, Nozaki N, Ishimi Y. Interaction of human MCM2-7 proteins with TIM, TIPIN and Rb. J Biochem 2010;147:917–27.

[41] de Lange T, de Lange T. Shelterin-Mediated Telomere Protection. Annu Rev Genet 2018;52:223–47. https://doi.org/10.1146/annurev-genet-032918-021921.

[42] Isono M, Niimi A, Oike T, Hagiwara Y, Sato H, Sekine R, et al. BRCA1 Directs the Repair Pathway to Homologous Recombination by Promoting 53BP1 Dephosphorylation. Cell Rep 2017;18:520–32.

[43] Madhukar NS, Elemento O, Pandey G. Prediction of Genetic Interactions Using Machine Learning and Network Properties. Front Bioeng Biotechnol 2015;3:172.