# Predicting phenotypic effects of gene perturbations in *C. elegans* using an integrated network model

**Karsten Borgwardt**

**Summary**

**Predicting the phenotype of an organism from its genotype is a central question in genetics. Most importantly, we would like to find out if the perturbation of a single gene may be the cause of a disease. However, our current ability to predict the phenotypic effects of perturbations of individual genes is limited. Network models of genes are one tool for tackling this problem. In a recent study, (Lee et al.) it has been shown that network models covering the majority of genes of an organism can be used for accurately predicting phenotypic effects of gene perturbations in multicellular organisms. *BioEssays* 30:707–710, 2008. © 2008 Wiley Periodicals, Inc.**

## Phenotype prediction

Clarifying the link between phenotype and genotype of an orgnanism is at the core of research in genetics. As genome sequences of species and, recently, individuals[2] continue to be established, an important research question over coming years will be to link variations in individual genes to phenotypic effects, most interestingly, genes that are involved in disease.

How to accomplish accurate predictions of phenotypic effects of gene perturbation, however, is an open research question. One approach is to represent the genes of an organism and their functional relationships in terms of a network model and to predict phenotypic effects using this model. This is the approach pursued by Lee et al. in their recent study on phenotype prediction in *C. elegans*.[1]

## Network model

What does such a network model look like? A network, or more mathematically, a graph, models objects and their relationships via nodes and edges. In the gene networks used for phenotype prediction here, each gene is modeled by a node, and edges indicate functional similarity between these genes. Each edge is assigned a probability score, indicating the probability that these two genes are involved in the same biological process. The higher this score, the more likely a functional relationship between these genes is.

Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK.
E-mail: kmb51@cam.ac.uk

Funding agencies: K.B. is funded by NIH grant GM063208-05 "Tools and Data Resources in Support of Structural Genomics."
DOI 10.1002/bies.20783
Published online in Wiley InterScience (www.interscience.wiley.com).

This functional relationship of genes can be defined in a multitude of ways: genes may be functionally similar if they have similar gene expression profiles, if there are known physical or genetic interactions between them or their orthologs in other model organisms, if there exist literature-mined associations between them, or if they are known to be functionally associated, co-inherited or related to the same operon in yeast, bacterial or archael homologs of *C. elegans* genes. Rather than choosing one of these definitions, Lee et al. developed an *integrated* network model from diverse datasets on all these types of relationships between genes.

## Data integration

Why is it promising to integrate several datasets? Experimental datasets are usually neither complete nor noise- and error-free. By integrating multiple datasets, one hopes to uncover reliable information—which is confirmed by several datasets—and to detect missing information—which is present in some datasets, but absent in others.

This process of data integration is highly non-trivial. On the one hand, simply taking the union of all datasets, i.e. creating a model that contains all edges that are present in *at least one* dataset, might lead to an error-prone network with poor predictive ability. On the other hand, only considering edges that are confirmed by *all* datasets, will produce a network model with low coverage and few edges, ignoring relevant information.

Lee et al. tackle this problem in two steps: First, they generate quality scores for each dataset. These quality scores measure whether the links in this dataset reflect whether genes share biological functions in terms of Gene Ontology (GO) annotations.[3] Second, they then integrate the datasets into a single network, referred to as Wormnet v1, using weights that are correlated with the quality scores. Wormnet covers 82% (16,113 genes) of all genes in *C. elegans*, whereas earlier studies did not include more than 20% of *C. elegans* genes.[4–7]

### Contributions of Lee et al.

The central contribution of this study by Lee et al. is to show that this single integrated gene network model allows for accurate phenotype prediction, even:

- for multicellular organisms,
- when considering whole-organism networks rather than special subgroups of genes,
- for specific tissues and developmental stages, despite using a single general network.

While single network models had been used in yeast before[8–10] and specialised subnetworks in other species,[4] Lee et al. are the first to demonstrate the predictive power of a single comprehensive gene network in a multicellular organism.

### Insights into network structure

In their study, Lee et al. employ this network model to explore the link between network topology and the essentiality of a gene, to predict phenotypic effects of perturbing individual genes, and to find genes that are part of a biological pathway.

### Gene connectivity and essentiality

First, they establish that the essentiality of a gene, i.e. its importance for the viability of the worm, and its connectivity, i.e. its number of neighbours within the network, are correlated. Hence the more neighbours a gene has, the more likely it is that its perturbation may have a lethal effect on the worm.

This correlation has been reported for protein interaction networks in yeast before,[11] but there were doubts whether it might hold for animal networks as well.[12] Lee et al. find a strong correlation between connectivity and essentiality of genes in their network model of *C. elegans* genes. They ensure that this finding is not an artefact of yeast-derived information being used in their network model by removing all yeast-based information from Wormnet. Even after this removal, essentiality and viability are found to be correlated.

To check whether this correlation might be present in mammals as well, they derive a subnetwork of Wormnet for genes with known orthologs in mouse. The correlation between connectivity and lethality can again be observed in this mouse network, indicating that this correlation might represent a fundamental principle of biology that is conserved across species.

### Prediction of RNAi phenotypes

Second, the authors assess the power of the network model to predict the loss-of-function phenotype of single gene perturbations induced by RNA interference. RNA interference (RNAi) is a process in which double-stranded RNA inhibits the expression of specific genes.[13]

Lee et al. examine 43 different loss-of-function phenotypes, as established by genome-wide RNAi screens. For each loss-of-function phenotype, a reference set of genes ('seed set') is known that exhibits this phenotype when perturbed. All genes in the network (including the genes in the reference set) are then ranked based on the number and strength of their links to this reference set (see Figure 1). The key idea is that a gene that is linked to many genes from the reference set might exhibit the same loss-of-function phenotype.

Using this ranking, 29 of the 43 phenotypes can be predicted with high accuracy, another 10 with an accuracy which is better than random.

These phenotypes that can be accurately predicted span a wide variety of cellular, developmental and physiological processes, and include phenotypes that specifically affect certain tissues or certain developmental stages of the worm. Hence this single network model of *C. elegans* genes shows a remarkably general ability for accurate phenotype predictions.

Only on 4 out of 43 phenotypes, the network model does not yield results that are better than random guessing. The failure on this minority of datasets might be caused by noise in the RNAi screens, by definitions of phenotypes that are not specific enough, or simply by information that is incomplete or missing in Wormnet.

### 'Network-guided screening'

Third, the network model can be used to identify genes that are involved in a particular biological pathway. The key idea is the same as for phenotype prediction: Find those genes in the network that are tightly linked to a set of genes that form a

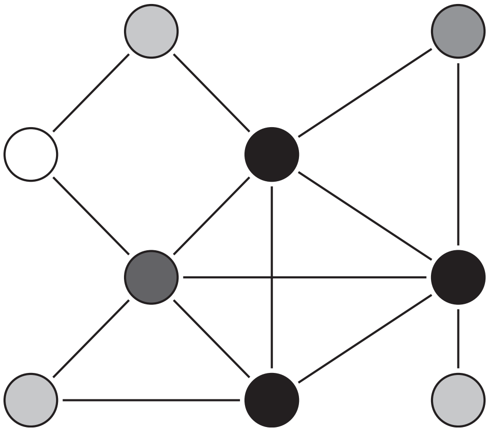

**Figure 1.** Schematic illustration of Lee et al.'s approach to phenotype prediction from network models. The reference set of genes with known common loss-of-function phenotype is represented by black nodes. The other genes in the network are ranked based on their connectivity to this reference set (the darker the node, the stronger its connectivity to the reference set): top-ranked nodes are predicted to exert the same loss-of-function phenotype as the reference set (All connections are of identical strength in this simple example).

biological pathway. These close 'neighbours' are then predicted to be members of the same pathway—an approach that Lee et al. refer to as 'network-guided screening'.

Lee et al. employ this approach to identify genes involved in the retinoblastoma tumor suppressor pathway and experimentally verified their predictions using RNAi screens inhibiting the candidate genes. The predictions of pathway membership based on Wormnet were up to 21-fold better than predictions based on random selection of nodes from the network.

They also experimentally confirmed a predicted connection between the dystrophin associated protein complex (DAPC) and the EGF signaling pathway in *C. elegans*. In Wormnet, the DAPC complex and the EGF signaling pathway are connected by three links, hinting at a functional relationship.

These three sets of experiments all indicate that this single network model of *C.elegans* enables insights into network structure, reaching from global topological phenomena, such as the correlation between connectivity and essentiality, to specific local phenomena, such as the link between the DAPC and the EGF signaling pathway.

## Next goal: a human network model

While Lee et al. are taking the step from yeast to worm in this study, they clearly point at the next natural goal: To develop a similar network model for human genes.

What lessons can be learnt from the development of Wormnet for the design of a human network model? First, the types of data used in Wormnet are already available for humans. Hence data availability is not an obstacle. Second, Wormnet is able to predict phenotypic effects on inidividual tissues, even though few of the data used for creating Wormnet are based on tissue-specific measurements. A similar network for humans genes might not require these tissue-specific data either.

Despite these promising results, what major obstacles might occur along the way? A potential problem of a human equivalent of Wormnet might be that it is based on similarity search: Wormnet can only predict the loss-of-function phenotype of a gene if other genes with exactly this phenotype are known. For many human diseases, often little is known about the genes involved. Still, Lee et al. found evidence that, even when only very few (less than six) genes of a particular phenotype are known, Wormnet is able to generate accurate phenotype predictions. This indicates that studies of human diseases, where only a few disease-related genes are usually known beforehand, might still be a realistic goal.

## Algorithmic challenges

In this study, Lee et al. have not only demonstrated the predictive power of a single gene network. They have also shown that putting enormous effort in data collection, preprocessing and integrating allows the generation of a network model that yields accurate phenotype predictions.

For machine learning in bioinformatics, an interesting challenge will be if a comparable effort in designing refined algorithms for data integration and prediction can further enhance Wormnet's prediction accuracy. To name a few specfic examples, can we improve phenotype predictions:

- by learning optimal weights for the different types of information that are integrated into Wormnet?
- by using algorithms that take the whole structure of the network model into account, not just individual links between genes?
- by using graphical models that allow to model the dependencies between genes explicitly, that is how silencing one gene affects other genes in the network?
- by explicitly predicting the functional role of a gene that is perturbed, rather than its general loss-of-function phenotype?

For the latter two questions of developing an even more refined gene network that allows for more specific function predictions, joint efforts of experimental groups and algorithmic groups will be invaluable. We feel that, if the generation of the data and algorithms for developing these future network models accompany each other, our ability to predict phenotypic effects of gene perturbations will significantly improve in the near future.

To conclude, Lee et al. have developed a single genome-wide network model for predicting phenotypic effects of gene silencing via RNAi in *C. elegans*. For years to come, phenotype prediction will pose several exciting research challenges at the interface of bioinformatics, systems biology and machine learning.

## Acknowledgements

K.B. is funded by NIH grant GM063208-05 ''Tools and Data Resources in Support of Structural Genomics.''

## References

1. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. 2008. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genetics 40:181–188.
2. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The Diploid Genome Sequence of an Individual Human. PLoS Biology 5:e254 EP.
3. Harris MA, Clark J, Ireland A, Lomaz J, Ashburner M, et al. 2004. The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.
4. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, et al. 2005. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. Nature 436:861–865.
5. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, et al. 2004. A map of the interactome network of the metazoan *C. elegans*. Science 303:540–543.
6. Stuart JM, Segal E, Koller D, Kim SK. 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302:249–255.

7. Zhong W, Sternberg PW. 2006. Genome-Wide Prediction of C. elegans Genetic Interactions. Science 311:1481–1484.
8. Lee I, Date SV, Adai AT, Marcotte EM. 2004. A probabilistic functional network of yeast genes. Science (New York, N.Y.) 306:1555–1558.
9. McGary KL, Lee I, Marcotte EM. 2007. Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. Genome Biology 8:R258.
10. Lee I, Li Z, Marcotte EM. 2007. An improved, bias-reduced probabilistic functional gene network of Baker's yeast, *Saccharomyces cerevisiae*. PLoS ONE 2:e988.
11. Jeong H, Mason SP, Barabsi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.
12. Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, et al. 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285–293.
13. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811.