Minireview

Wormnet: a crystal ball for *Caenorhabditis elegans*

Stephen E Von Stetina and Susan E Mango

Address: Huntsman Cancer Institute, University of Utah, Circle of Hope, Salt Lake City, Utah 84112, USA.

Correspondence: Stephen Von Stetina. Email: stephen.vonstetina@hci.utah.edu. Susan E Mango. Email: susan.mango@hci.utah.edu

Published: 2 June 2008

*Genome **Biology*** 2008, **9**:226 (doi:10.1186/gb-2008-9-6-226)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/6/226

Abstract

An integrated gene network for *Caenorhabditis elegans* using data from multiple genome-wide screens encompasses most protein-coding genes and can accurately predict their phenotypes.

'How-to' books tell us that networking is critical to get ahead in business and in life. Networks are also becoming increasingly important in biology, as we grapple with whole genome sequences. The traditional approach - to study one gene at a time and place it in a linear pathway with a defined biological role - falters when faced with thousands of genes, and genes with roles in two or more processes. In a recent paper in *Nature Genetics*, Lee *et al*. [1] confront these challenges by constructing a probabilistic network for *Caenorhabditis elegans*. This network differs from those of previous studies in that it captures most of the protein-coding genes in the *C. elegans* genome (82%), and it can use groups of genes to search for interacting loci.

Since its genome was completed a decade ago, *C. elegans* has emerged as a powerhouse for genome-wide analyses [2]. Large-scale surveys for RNA-interference (RNAi)-induced phenotypes [3,4], RNA expression [5], protein-protein interactions (Interactome [6]), and protein-DNA binding [7] have generated a wealth of information about approximately 20,000 *C. elegans* genes. Our current challenge is to integrate these data into a coherent picture. In an early study, Kim *et al*. [5] combined data from multiple *C. elegans* microarray experiments, as well as those from *Drosophila,* yeast and humans, to find genes that were co-regulated across species. The authors made use of the 'guilt-by-association' concept to ask if genes that were coexpressed over many different conditions had similar functions. Gunsalus and co-workers [8] combined coexpression data, Interactome data and phenotypic analyses to predict the molecular machines that drive early embryogenesis. The Sternberg lab took this idea one step further, expanding predictions for all stages of life. Zhong and Sternberg [9] combined coexpression data, interactome predictions and genetic or protein interactions from worms or their orthologous genes and proteins in flies and yeast. The data were weighted according to their dependability and integrated into a Bayesian network with over 18,000 interactions for 2,254 genes, or around 11% of the predicted worm proteome. More recently, the Vidal lab developed an automated method to classify post-embryonic expression patterns of promoter-green fluorescent protein (GFP) reporters [10]. They combined anatomical data with the interactome dataset to weed out potential false positives for genes that were not expressed in the same tissue. These studies laid the foundation for integrated networks, but failed to capture the majority of protein-coding genes in their searches.

Wormnet is a probabilistic network for *C. elegans*

Now, Lee and colleagues [1] have assembled diverse data from *C. elegans* large-scale analyses to build a probabilistic network, dubbed Wormnet. Much of the information used by Zhong and Sternberg was also used by Lee *et al*. Additional information in Wormnet was derived from the following sources: gene interactions inferred from co-citation analysis, with the assumption that gene pairs that are co-cited in abstracts more than the random expectation are likely to be functionally linked; 'associalogs', which represent physical or genetic interaction data from other species mapped onto their *C. elegans* orthologs, as determined by INPARANOID [11,12]; and phylogenetic and gene neighbor analysis using 117 bacterial genomes (Figure 1). The current study excluded Gene Ontology (GO) terms and

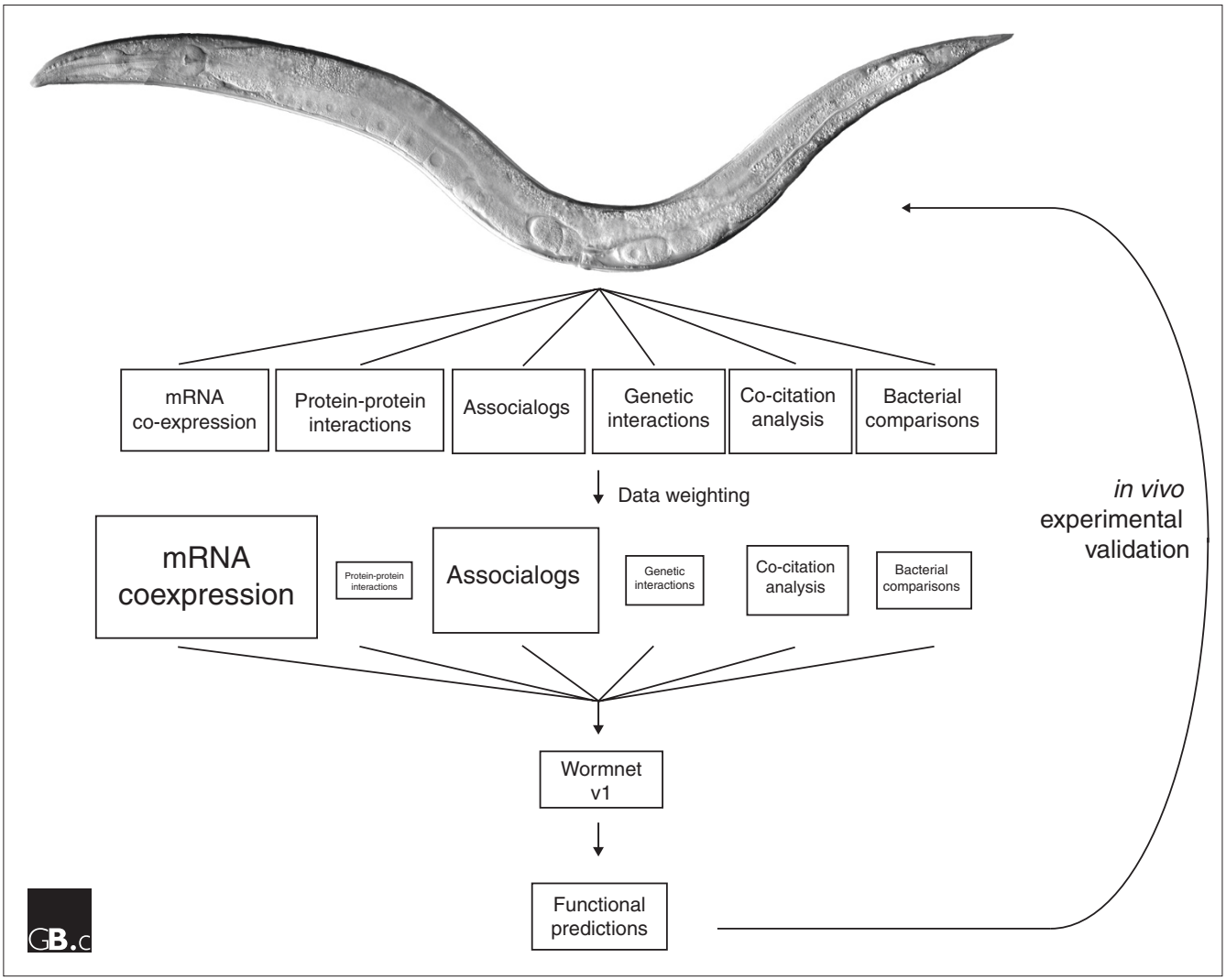

Figure 1
The conceptual framework of Wormnet. Wormnet was generated from large-scale studies of *C. elegans* biology (boxes). Data for each gene were weighted according to their accuracy, which is diagrammed here as differently sized boxes. Wormnet can be used to make predictions about gene function, which can be rapidly tested *in vivo*. 'Associalogs' refers to physical or genetic interaction data from yeast, flies, and humans mapped to their nematode orthologs.

RNAi phenotypic data, opting to use this information for data weighting and validation, respectively.

The assembled data were weighted and integrated into a comprehensive network using methodology the Marcotte lab had optimized previously for yeast [13]. Briefly, Lee *et al*. [1] determined how well each dataset (that is, the coexpression dataset, the Interactome dataset, and so on) predicted a meaningful linkage between genes known to share biological functions, based on GO annotations. The weighting of the datasets was performed for each individual link to provide more sensitive scoring. From this analysis, a log likelihood score (LLS) was calculated, which estimated the probability that two genes were linked in a meaningful way. Each dataset was given different weights depending on how well it estimated functional linkages, so that higher-quality data 'counted' more than the lower-quality data when it was incorporated into Wormnet. The power of this approach is that LLS scores are additive, allowing easy integration of different data points using Bayesian statistics. In addition, this flexibility permits addition of future data as they become available. Thus, assembling and weighting diverse groups of data, even poor-quality data, can accumulate evidence for a functional interaction between genes.

Using these criteria, Wormnet v1 established 384,700 interactions among 16,113 genes (approximately 80% of the proteome), a four- to eightfold increase in coverage over previous studies [14]. The authors trimmed this network to produce a higher-confidence dataset using an empirically

defined LLS score cutoff, defined by their previous work in yeast [13]. This analysis generated the core Wormnet group, consisting of 113,928 linkages between 12,357 genes (approximately 63% of the proteome). Even this trimmed version of Wormnet constitutes a threefold increase in proteome coverage over previous studies with worms. Intriguingly, 83,946 linkages in the core database had never been noted elsewhere (for example, in the GO database or in the literature).

Testing Wormnet

To validate Wormnet, the authors queried the core database in four ways. First, they determined if Wormnet could predict essential genes. Interactome data from one- and two-hybrid screens have revealed that proteins with many interacting partners are likely to be essential [6,7,15]. This is called the lethality-centrality rule, and it also holds for Wormnet. Lee and colleagues [1] observed a good correlation between genes with many Wormnet linkages and the likelihood those genes would be essential, based on data derived from a genome-wide RNAi screen [3]. The RNAi dataset was not used to build Wormnet and therefore served as an independent test group. The authors extended their analysis to focus on the subset of *C. elegans* genes with mouse orthologs. They discovered that Wormnet could accurately predict genes with lethal phenotypes for mice as well as worms.

Second, the authors determined if genes connected to each other by Wormnet were associated with similar phenotypes. They examined 43 genome-wide RNAi screens that were focused on a particular phenotype such as 'increased lifespan' or 'growth defective.' Lee *et al.* found a strong correlation between linked genes in Wormnet and related phenotypes for 29 of the 43 RNAi screens, with another 10 screens having reasonable linkages. Thus, genes connected by Wormnet were likely to have similar phenotypes and, by extension, roles in similar cellular or developmental processes. This relationship, however, does not predict similar biochemical functions. For example, a pair of linked genes might reflect one activator and one repressor, both acting in a common pathway.

Next, Lee *et al.* examined whether Wormnet could predict specific functions for unstudied genes, based on their linkages to known genes. They chose two pathways implicated in human disease. First, they surveyed Wormnet for genes that might function in the retinoblastoma (Rb) tumor suppressor pathway. In *C. elegans*, the Rb pathway is best understood for its role in the developing vulva, which is the egg-laying apparatus for the worm. Previous studies had identified six genes that could suppress Rb-associated vulval phenotypes [16,17]. The authors used these six genes as a seed to search for interacting loci, and identified 62 genes from the core Wormnet dataset. Using RNAi, they tested 50 of these genes and found 10 that produced scoreable suppression for vulval development, a hit rate of 20%. This was a significantly higher frequency compared with a recent genome-wide screen, which identified suppressors at a rate of around 0.4% [18]. Thus, Wormnet could pinpoint a set of candidates to test, and it improved the likelihood of success by orders of magnitude over an unbiased screen. However, neither the genome-wide nor the Wormnet screen was perfect: more than 70% of the suppressors discovered by Cui and co-workers [18] were missed by Wormnet, and conversely 38% of the Wormnet suppressors were not found by Cui *et al.* Some genes missed by Wormnet reflect pathways not represented by the six seed genes. Nevertheless, Wormnet successfully identified components for each of the chromatin regulatory complexes that were also discovered by Cui *et al.*

For the fourth test, Lee *et al.* examined an interaction predicted by Wormnet between the dystrobrevin-associated protein complex (DAPC) and the epidermal growth factor (EGF)-Ras-MAP kinase (MAPK) signaling pathway. DAPC components are primarily expressed in muscle cells, and mutation of several DAPC genes are linked to muscular dystrophies [19]. The EGF pathway is perturbed in many human cancers, but in *C. elegans* it is critical for cell-fate specification. RNAi of three DAPC genes strongly suppressed the cell-fate phenotypes associated with activated Ras, suggesting that DAPC augments EGF-Ras-MAPK signaling. As the authors point out, this relationship may be conserved in vertebrates [20], suggesting novel therapeutic targets for muscular dystrophies.

Where do we go from here?

What does the future hold for Wormnet? Adding new data will extend and refine the Wormnet database. The Model Organism Encyclopedia of DNA Elements (modENCODE) [21] is an effort to uncover functional elements in the fly and worm genomes, including additional protein coding sequences, noncoding RNAs and *cis*-regulatory regions. These important elements will aid the prediction machinery, for example, by increasing the proteome coverage of Wormnet from its current level of 80%. Inclusion of noncoding RNAs such as microRNAs [22] could add a whole new twist to understanding regulatory pathways. In addition, the current version of Wormnet does not rely on explicit spatial or temporal expression data. With the advances in tissue-specific profiling [23-27], future versions of Wormnet could allow researchers to restrict their database searches to the subset of genes active in a tissue of interest (A Fraser, personal communication). This approach may reduce the number of false positives identified in a search. With an almost exponential increase in genome-wide datasets expected in the coming years, it is conceivable that Wormnet will soon cover the entire worm proteome and greatly aid in the discovery of gene function.

In summary, Lee and colleagues [1] have built an integrated database that can uncover genetic linkages between genes for *C. elegans* and probably also for mammals. One big payoff of this study is the enhanced predictive power. Wormnet can detect interactions not only between components of stable complexes (for example, the proteasome), but also factors associated with dynamic processes, such as cell signaling. Put another way, Wormnet describes the possible linkages associated with a gene, only some of which will be active at any particular time or place. This may enable Wormnet to uncover links for proteins with diverse functions. Many proteins participate in more than one process - consider the roles of β-catenin in transcription versus cell adhesion [28], or of the GTPase Ran in nuclear trafficking versus mitotic spindle assembly [29]. Probabilistic networks are capable of building gene linkages that represent multiple biological roles, rather than placing genes in traditional linear pathways. Wormnet provides an excellent resource for the field of *C. elegans* biology, and the principles set forth by these studies can also be applied to more complex organisms.

Acknowledgments

We thank David Miller for providing the light micrograph for Figure 1, Dean Billheimer, Andy Fraser and Min Han for discussion, and Alex Schier for reading the manuscript. SEM was supported by NIH R01 GM056264, the Huntsman Cancer Institute and the Department of Oncological Sciences.

References

1. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM: **A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*.** *Nat Genet* 2008, **40:**181-188.
2. **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
3. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J: **Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi.** *Nature* 2003, **421:**231-237.
4. Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, Hewitson M, Holz C, Khan M, Lazik S, Martin C, Nitzsche B, Ruer M, Stamford J, Winzi M, Heinkel R, Röder M, Finell J, Häntsch H, Jones SJ, Jones M, Piano F, Gunsalus KC, Oegema K, Gönczy P, Coulson A, *et al.*: **Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*.** *Nature* 2005, **434:**462-469.
5. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302:**249-255.
6. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, *et al.*: **A map of the interactome network of the metazoan *C. elegans*.** *Science* 2004, **303:**540-543.
7. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ: **A gene-centered *C. elegans* protein-DNA interaction network.** *Cell* 2006, **125:**1193-1205.
8. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sönnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F: **Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis.** *Nature* 2005, **436:**861-865.
9. Zhong W, Sternberg PW: **Genome-wide prediction of *C. elegans* genetic interactions.** *Science* 2006, **311:**1481-1484.
10. Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A, Carvunis AR, Pulak R, Shingles J, Reece-Hoyes J, Hunt-Newbury R, Viveiros R, Mohler WA, Tasan M, Roth FP, Le Peuch C, Hope IA, Johnsen R, Moerman DG, Barabási AL, Baillie D, Vidal M: **Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*.** *Nat Biotechnol* 2007, **25:**663-668.
11. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL: **InParanoid 6: eukaryotic ortholog clusters with inparalogs.** *Nucleic Acids Res* 2008, **36(Database issue):**D263-D266.
12. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052.
13. Lee I, Li Z, Marcotte EM: **An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*.** *PLoS ONE* 2007, **2:**e988.
14. **Wormnet** [http://www.functionalnet.org/wormnet]
15. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
16. Wang D, Kennedy S, Conte D Jr, Kim JK, Gabel HW, Kamath RS, Mello CC, Ruvkun G: **Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants.** *Nature* 2005, **436:**593-597.
17. Lehner B, Calixto A, Crombie C, Tischler J, Fortunato A, Chalfie M, Fraser AG: **Loss of LIN-35, the *Caenorhabditis elegans* ortholog of the tumor suppressor p105Rb, results in enhanced RNA interference.** *Genome Biol* 2006, **7:**R4.
18. Cui M, Kim EB, Han M: **Diverse chromatin remodeling genes antagonize the Rb-involved SynMuv pathways in *C. elegans*.** *PLoS Genet* 2006, **2:**e74.
19. Durbeej M, Campbell KP: **Muscular dystrophies involving the dystrophin-glycoprotein complex: an overview of current mouse models.** *Curr Opin Genet Dev* 2002, **12:**349-361.
20. Chockalingam PS, Cholera R, Oak SA, Zheng Y, Jarrett HW, Thomason DB: **Dystrophin-glycoprotein complex and Ras and Rho GTPase signaling are altered in muscle atrophy.** *Am J Physiol* 2002, **283:**C500-C511.
21. **The modENCODE project** [http://www.modencode.org]
22. Kato M, Slack FJ: **microRNAs: small molecules with big roles - *C. elegans* to human cancer**. *Biol Cell* 2008, **100:**71-81.
23. Gaudet J, Mango SE: **Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4.** *Science* 2002, **295:**821-825.
24. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL, Chan S, Dube N, Fang L, Goszczynski B, Ha E, Halfnight E, Hollebakken R, Huang P, Hung K, Jensen V, Jones SJ, Kai H, Li D, Mah A, Marra M, McGhee J, Newbury R, Pouzyrev A, Riddle DL, Sonnhammer E, Tian H, Tu D, Tyson JR, Vatcher G, Warner A, *et al.*: **Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*.** *Cold Spring Harbor Symp Quant Biol* 2003, **68:**159-169.
25. Reinke V: **Functional exploration of the *C. elegans* genome using DNA microarrays.** *Nat Genet* 2002, **32(Suppl):**541-546.
26. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*.** *Nature* 2002, **418:**975-979.
27. Von Stetina SE, Watson JD, Fox RM, Olszewski KL, Spencer WC, Roy PJ, Miller DM 3rd: **Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the *C. elegans* nervous system.** *Genome Biol* 2007, **8:**R135.
28. Clevers H: **Wnt/beta-catenin signaling in development and disease.** *Cell* 2006, **127:**469-480.
29. Joseph J: **Ran at a glance.** *J Cell Sci* 2006, **119:**3481-3484.