Network-guided genetic screening: building, testing and using gene networks to predict gene function

Ben Lehner and Insuk Lee

Advance Access publication date 29 April 2008

Abstract

A challenge facing nearly all biologists is to identify the complete set of genes that are important for a process or disease. This applies to scientists investigating fundamental pathways in model organisms, but also to clinicians trying to understand human disease. There are many different types of experimental data that can be used to predict the genes that are important for a process, but these data are normally dispersed across numerous publications and databases, and are of varying and unknown quality. Integrated functional gene networks aim to gather functional information from all of these data into a single intuitive graph model that can be used to predict gene functions. In this approach, the ability of each data set to predict functional associations between genes is first measured using a standard benchmark, and then the scored predictions by each data set are combined. The resulting integrated probabilistic gene network can be used by all researchers to predict gene function, with much greater coverage and accuracy than any individual data set. In this review, we discuss how such integrated gene networks are constructed, how their predictive power for gene function can be tested, and how experimental biologists can use these networks to guide their research. We pay particular attention to such networks constructed for *Caenorhabditis elegans*, because in this complex multicellular model system functional predictions for genes can be rapidly tested *in vivo* using RNAi. The approach is, however, widely applicable to any system, and might soon be a common method used to dissect the genetics of human complex diseases.

Keywords: *gene networks; data integration; systems biology;* C. elegans; *network-guided genetic screening*

WHAT IS A FUNCTIONAL GENE NETWORK?

A functional gene network is a network that connects two genes that share a common function [1–8]. That is, two genes (network nodes) are connected (network edge) if they participate in a common biological process or pathway. The edges in a functional gene network may, or may not, represent direct physical interactions between gene products. A functional gene network is therefore a slightly more abstracted representation of a biological system than networks based on direct physical interactions between gene products such as protein–protein or protein–DNA interaction networks. Indeed, this is precisely the point of constructing such a functional network—it allows many disparate types of data to be integrated and represented by a single graph model [9], rather than by the superposition of many different networks, each representing a different aspect of the relationship between molecules [10].

In this review, we do not specifically discuss the various types of biological networks that have been (and will be) constructed for model organisms. This topic has been extensively reviewed before [10, 11]. Rather, we aim to discuss the reasoning behind integrating these diverse data-types into a single

Corresponding authors. Ben Lehner, EMBL-CRG Systems Biology Research Unit and Institució Catalana de Recerca i Estudis Avançats (ICREA), Centre for Genomic Regulation, UPF, Barcelona 08003, Spain. E-mail: ben.lehner@crg.es or Insuk Lee, Department of Biotechnology, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul 120-749, Korea. E-mail: insuklee@yonsei.ac.kr

Ben Lehner is a group leader at the EMBL-CRG Systems Biology Unit and the Institució Catalana de Recerca i Estudis Avançats in Barcelona, Spain. Previously, he was a postdoctoral fellow in the Fraser Lab at The Wellcome Trust Sanger Institute, UK.

Insuk Lee was a research associate in the Center for Systems and Synthetic Biology of the University of Texas at Austin, USA. He is currently an assistant professor in the Department of Biotechnology at Yonsei University in Seoul, Korea.

'functional' network, and how such a functional network can be used to advance research.

The main use of a functional gene network is to predict novel gene functions and loss-of-function phenotypes. For example, if gene A is connected to gene B in a functional network, and we know that gene B is required for function X, then using a 'guilt by association' argument we can predict that gene A is also likely to be involved in function X. If gene A is also connected to genes C, D and E that are also involved in function X, then gene A is even more likely to be a component of function X.

One of the most important motivations behind functional gene networks is that they gather together many different data-types into a single, and easily accessible resource. Moreover, by measuring the accuracy and coverage of each of these data sets using a common measure of whether two genes participate in the same pathway or process, integrated networks allow the utility of each of these data sets to be directly contrasted and compared. This is useful both for the bench scientist unsure about the quality of the data contained in each data set, and also for the community as a whole because it highlights areas of data paucity.

A second major advantage of integrated networks is that they are 'more than the sum of their parts' when it comes to predicting gene function or loss-of-function phenotypes. Many individual data sets are largely complementary, so by combining data sets together the resulting integrated network can cover both more genes and more interactions. However, the increase in coverage that results from combining data does not necessarily reduce the accuracy of networks. If the accuracy of each data set is measured, and only the high-quality part of each data set is used in the final network, then the integrated network can still maintain high accuracy. Indeed, because of this benchmarking and integration, the accuracy of an integrated network will normally be higher than any individual data set.

HOW DO YOU BUILD A FUNCTIONAL GENE NETWORK?

Overview

The aim of an integrated gene network is to combine many different types of biological data together into a single network of interactions between genes. The simplest method to do this is to sum together the interactions between genes found in multiple different data sets. However, each of these individual data sets will be of very different qualities, so this simple summation normally results in a network that, although seemingly having high coverage, is of low (or unknown) accuracy [12].

One simple method to avoid the reduction in accuracy that results from combining many low-quality data sets is to only consider the 'overlap' between data sets as a final higher confidence network [12–16]. For example, genes A and B may only be connected in a final network if they interact in more than one of the constituent data sets. This 'overlap' approach does indeed increase the accuracy of the final network. However, it also reduces the coverage very dramatically—interactions that are only supported by one line of evidence may be of very high confidence but these interactions will always be excluded from a final network using this approach.

The shortcomings of both the simple summation approach and the overlap approach have led to the development of more sophisticated methods for constructing integrated networks [3, 7, 17–19]. In these approaches, the individual lines of evidence are combined in a probabilistic framework. That is, the accuracy of each individual data set is first measured using a common benchmark, and then the evidence from each individual data set is combined according to the measured accuracy of each individual data set. The final network is thus a probabilistic network of functional interactions between genes, with each interaction associated with a confidence score.

Data sets that can be used to construct a gene network

There are many different types of experimental data that can be used to infer functional association between genes. These include (references are given for *Saccharomyces cerevisiae* and *Caenorhabditis elegans* data sets only) genome-scale protein–protein interactions from various experimental detection methods such as yeast two-hybrid [20–23] and affinity purification followed by mass spectrometry analysis [24–26]; the tendency of mutations in genes that act in the same process to produce non-additive phenotypes [27–33]; and the tendency of these genes to be co-expressed across conditions [34, 35]. We also can infer many functional gene associations by computational analysis of genome context, for example, the co-inheritance pattern of genes across genomes (phylogenetic profiling) [36–38] and the

Table I: Some data types that can be used to predict functional linkages between genes

Data type	Description
Protein interactions	Gene products that physically interact are likely to share a common function.
Protein complexes	Components of a protein complex are likely to share a common function.
Genetic interactions	If mutations in two genes result in synthetic phenotypic consequences, then these two genes are likely to share a common function.
Gene co-expression	Genes that are co-expressed across conditions are likely to share a common function.
Co-regulation	Genes regulated by the same upstream regulators are likely to share a common function.
Gene neighbourhoods	Genes whose orthologues are adjacent in bacterial genomes (are likely to be transcribed in the same operon) are likely to share a common function.
Gene fusions	Genes that are fused into a single open reading frame in a different species are likely to functionally interact.
Co-inheritance of genes across species	Genes that are either both present, or are both absent in many genomes are likely to share a common function.
Co-citation	Genes that are described in the same publications are likely to share a common function.
'Associalogues'	Genes whose orthologues in one species share a function are also likely to share a common function in a second species of interest.

proximal chromosomal location of bacterial orthologues (gene neighbouring that measures the likelihood of bacterial orthologues being in the same operon) [34, 39, 40]. Importantly, a network can also use interactions transferred from many different species as interactions between orthologous genes [41]. We provide a non-exhaustive list of these data types in Table 1, and as new experimental data sets and computational approaches are developed the number of data sets that can be used will increase. For example, signalling [42] and regulatory [43, 44] networks will also likely provide very useful data that can be integrated in future network releases.

Benchmarking data sets

One very important factor that affects the quality of a constructed gene network is the choice of benchmarking interactions, often referred to as 'gold-standard positive' (GSP) interactions—i.e. those that researchers consider very likely to be true, and 'gold-standard negative' (GSN) interactions—i.e. those interactions that researchers consider most likely not to occur. The benchmarking interactions must not be explicitly included into the network.

To build a set of GSP interactions requires an accurate measure of gene function. This in itself is not an easy task, both because most genes have many functions and also because the definition of gene function can be interpreted on many different levels. Pragmatically, to date, researchers have tended to use the Gene Ontology (GO) database of gene function annotations [45]. This database represents gene functions as a hierarchy of annotations, and at least those annotations derived from the literature should be of sufficiently high quality. Empirically, we have found that using the 'biological process' annotations from GO provides a useful measure of whether two genes participate in a common biological pathway or system. However, improvements in the coverage and accuracy of functional annotation databases should lead to significant improvements in the standard of integrated gene networks. GSP interactions can be constructed by pairing genes sharing at least one GO biological process term. GSN interactions can be generated by pairing genes that are both annotated with a GO biological process term, but do not share a term.

Having built a set of GSP and GSN interactions, it is next necessary to use these to measure the quality of each data set that is being considered for inclusion into an integrated gene network. One measure that has proved popular for this purpose is a likelihood ratio [3, 17]. These ratios, similar to the odds ratios used in human disease gene mapping, measure how well a data set connects genes that are known to share a function (interactions in the GSP set) compared to those that do not share a function (interactions in the gold-standard negative set). This measure is then normalized by the prior expectation based on picking random genes.

Integrating data sets

The advantage of using likelihood scores or similar measures of the accuracy of each data set is that it allows the different types of evidence to be integrated together weighted according to the confidence that each data set is making correct predictions (Figure 1). A popular method for

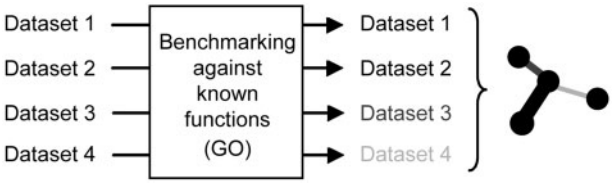

Figure 1: Constructing an integrated functional gene network. An integrated functional network connects genes by an interaction if the two genes are likely to share the same function [9]. Each functional or comparative genomic data set is first measured for its ability to connect genes that are known to share the same functional annotations, for example, using GO annotations [45]. The interactions predicted by each data set are then integrated together, but weighted according to how well each data set performs in the benchmarking [3]. In the example, data set 1 performs well and so interactions predicted by this data set receive high confidence scores, data set 4 performs very poorly and so interactions predicted by this data set receive low confidence scores. The final integrated network is therefore probabilistic with each interaction having an attached likelihood score (here dark thick lines represent higher confidence interactions).

integrating likelihood scores is to use a Bayesian framework [3, 7, 17–19], although other methods such as decision trees [46] and other machine-learning approaches may also become more widely used. If each line of evidence is assumed to be independent, then the data sets can be combined together as a simple sum (naïve Bayesian integration). However, in reality, many different biological data sets cannot be considered to be truly independent—for example, many different data sets measure whether two proteins interact using related experimental or computational methods. Therefore, modifications have been made to the naïve Bayesian integration whereby the likelihood scores are not combined as a simple sum, but each additional evidence is further penalized to account for any dependence between data sets (Box 1 and [3]). Alternatively, more complex, fully connected Bayesian integrations can be used, which accommodate for any correlated evidence sources [17].

Following integration of many different data sets, each edge in a final network has an associated score that represents the confidence that the interaction is true. That is, the final network is probabilistic, with some interactions of very high confidence and other interactions with lower confidence. The final network can be considered a 'family' of networks—by changing the level of confidence you are prepared to work with, you can use a smaller network where all the interactions are of very high confidence, or a large network with interactions of both high and lower confidence.

HOW DO YOU TEST A FUNCTIONAL GENE NETWORK?

Having constructed a gene network, it is important to test how good it is. The best way to do this is to use many new, independent experiments (see subsequently). However, in reality, it is not realistic to perform new experiments testing predictions for every pathway in an organism, so an alternative strategy is to use existing experimental data. The experimental data set used for testing a network should be large, diverse (covering many different systems in the organism) and it should, of course, not have been used either in the construction of the network or for its benchmarking. One good source of such data is reverse genetic data. In both yeast and *C. elegans*, genome-wide reverse genetic screens have now been performed for very many diverse processes (these screens are collated in [47] and [4]). In these screens, the effects of deleting or inhibiting the expression of nearly all genes have been assayed for a particular process. Provided that none of this data was used to construct or benchmark an integrated network, then the data from these screens can be used to test the predictive power of a network for many diverse processes.

The basic test is whether—the genes that affect a phenotype are more connected in a network to each other than to random genes? To put it another way, if we were given a subset of these genes, could we use the network to predict many of the additional genes that have been found to affect a process?

One way to test the ability of a network to correctly predict the genes that are associated with a process is to plot 'receiver operating characteristic' (ROC) curves, as summarized in Figure 2 [4, 6, 47]. In this approach, all genes in the genome are scored by summing the scores of all their connections to the seed genes. A gene with many high-confidence interactions to the seed genes will receive a high score, and a gene only connected by a single weak interaction will receive a very low score. Gene with no direct connection to the seed genes will receive a score of zero. This score is calculated for all genes including the seed genes themselves, and all the

Box I: Benchmarking and integrating data to construct a functional gene network

Biological data are heterogeneous and differ in quality and predictive power for identifying interacting genes (or genes sharing a function). To integrate many different data sets we need to evaluate them using a common reference set—this is called benchmarking the data. After this benchmarking using a common measure, the original data become standardized and thus can be easily integrated into a single model. How data sets can be best benchmarked and integrated is a field of active research, and further work in this area should lead to significant improvements in network coverage and accuracy.

The probability of two genes interacting (or sharing a function) can be conveniently scored with Bayesian statistics [3, 7, 17–19]. In this statistical framework, we measure how likely two genes are to interact in a set of trusted interactions (GSP interactions) if they interact in an experimental data set, compared to the random expectation of two genes interacting in the GSP data set. More formally, log likelihood score *(LLS)* can be calculated using the following equation:

$$LLS = \ln\left(\frac{P(L|D)/P(\sim L|D)}{P(L)/P(\sim L)}\right),$$

where $P(L|D)$ and $P(\sim L|D)$ are the frequencies of interactions in the GSP data set (L) and in the GSN data set (a set of gene pairs that do not interact) $(\sim L)$ *for genes that interact in* a given experimental data set (D). $P(L)$ and $P(\sim L)$ represent the prior random expectations (the total frequencies of positive and negative interactions, respectively).

For data sets in which each gene pair is associated with a continuous data-intrinsic score (e.g. correlation coefficients from co-expression data, mutual information from comparative genomic data), *LLS* scores can be calculated for bins containing equal numbers of rank-ordered gene pairs. These *LLS* scores and their corresponding data-intrinsic scores (the mean scores for a bin) can be used to calculate regression models, which are then used to map individual data-intrinsic scores to *LLS* scores in a continuous manner.

For integrating *LLS* scores from different data sets, we can simply take a sum of all LLS assuming complete independence among data sets. This naïve Bayesian integration is simple but not ideal, because correlation is frequent among the data sets to be integrated. One alternative data-integration method is to use a weighted sum method that is a variant of naïve Bayesian integration that takes into account correlations among the data sets. For example, the following formula with two free parameters (T, a LLS threshold that each interaction must exceed, and D, a parameter that represents the relative dependency between data sets) can be used to integrate data sets:

$$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \cdot i}, \quad \text{for all } L \geq T,$$

where L_0 represents the maximum *LLS* score for a given gene pair, and i is the rank order index of *LLS* scores, ranking gene pairs starting from the second-highest *LLS* with descending magnitude for all n remaining *LLS* scores. For integration, we consider only *LLS* scores above the threshold T, thereby excluding noisy low scoring linkages. The free parameter D ranges from 1 to $+\infty$, and is optimized to maximize overall performance (measured by the area under a recall-precision curve) of the integrated model. Low D-value indicates more independence among the data sets.

An implicit feature of such a training procedure using benchmarking is the introduction of free parameters. Introducing too many parameters may result in a trivial 'memorization' of the reference interactions rather than actual learning that will be predictive for many unknown interactions. One way to detect overtraining is to use only a subset of the reference data set for training, retaining the rest for testing the predictive power of a network. One common approach to use is k-fold cross-validation. Here the entire training set is randomly divided into k subsets with no overlap, and then the network is trained using $k-1$ subsets and tested using the unused subset. The average error rate between trained and tested scores from k repeats of validation can then be used to detect any overtraining. Another approach is 0.632 bootstrapping where each new training set contains on average 63.2% of original training set using sampling with replacement [61]. This method is particularly suited for learning with the small training sets.

genes in the genome are then ranked according to their score. To assess how well the seed genes are returned in preference to random genes, the proportion of seed genes recovered as you descend the list (sensitivity) is plotted against the proportion of non-seed genes recovered 1-specificity (Figure 2). In this way, ROC curves can be used to compare the recovery of 'true positives' (y-axis) and 'false

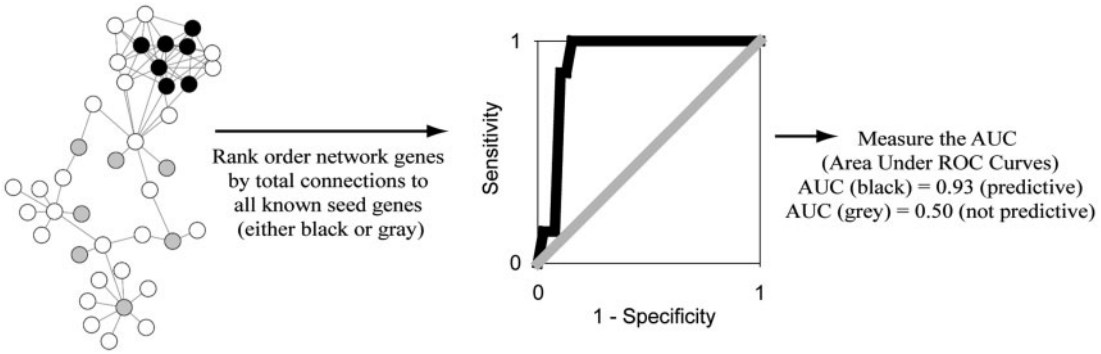

Figure 2: Scoring the predictive power of a network using ROC curves analysis. This schematic figure illustrates ROC curve analysis with examples of two contrasting sets of seed genes, one (black node) is highly predictive using the network, the other (grey node) is not predictive using the network. All genes (both seed and non-seed genes) in the network are scored by their total connections to a set of seed genes (either black or grey). Each curve plots the recovery of known 'seed' genes (sensitivity, *y*-axis) against the recovery of non-seed genes 1-specificity, *x*-axis as the list of genes ranked by their connectivity score to the seed genes is descended. Black seeds are well connected to each other in the network and scored higher than non-seed genes. Therefore, most of seed genes are recovered before non-seed genes, with the resulting ROC curve approaching left upper corner of the plot (black curve line). In contrast, grey seeds are not connected at all and score zero. Therefore, the recovery rate of grey seeds follows the random expectation (grey diagonal line). The behaviour of an ROC curve can be summarized by area under the ROC curve (AUC) value. Highly connected black seed genes result in a high AUC (0.93), but disconnected grey seed genes result in the AUC of random expectation (0.5).

positives' (*x*-axis). In a 'perfect' network, using data from a 'perfect' screen the network should return all true positives with no false positives, and the ROC curve should follow the *y*-axis and the area under the curve (AUC) would equal 1. In the case of a very poor network, the ROC curve would lie along the diagonal of the plot (true positives and false postives recovered with equal probability) and the AUC would equal 0.5. Therefore, ROC curves provide a visual assessment of the ability of a network to connect a set of genes, and the AUC values provide a numerical measure of this.

We illustrate how reverse genetic data and ROC curve analysis can be used to test a gene network in this way using the 'Wormnet' network for *C. elegans* [4]. This network is currently the most extensive network for *C. elegans* and can be searched with groups of genes and downloaded from http://www.functionalnet.org/wormnet. In *C. elegans*, there have been >40 phenotypes scored in genome-wide RNAi screens (reviewed in [48] and [49]). In Figure 3, ROC curves for this network are shown for each of 43 RNAi phenotypes. As can be seen in Figure 3A, for 29 of these phenotypes the genes are very clustered in the network. That is, their ROC curves lie to the top left of the plots and the genes with these phenotypes can be successfully predicted using Wormnet. For a further 10 phenotypes (Figure 3B) the associated genes are reasonably connected, but for four phenotypes the genes are no more connected than random genes (Figure 3C). The genes associated with these four phenotypes can therefore not be predicted using Wormnet. Also, apart from showing how well a network can be used to predict the genes associated with many different phenotypes, ROC curves can also be used to compare the predictive power of different networks [4].

The main conclusion from the ROC analysis applied to both *C. elegans* [4] and yeast [47] is that a single integrated network constructed using currently available data can be predictive for the vast majority of the loss-of-function phenotypes that have ever been tested in these species. Since these phenotypes range across a very wide range of processes and scales, we can conclude that a *single* integrated network can be predictive for many different aspects of the biology of both unicellular organisms and multicellular animals.

HOW CAN YOU USE A FUNCTIONAL GENE NETWORK?

We have described above how high-quality functional gene networks can be constructed using existing data. Several of these networks have

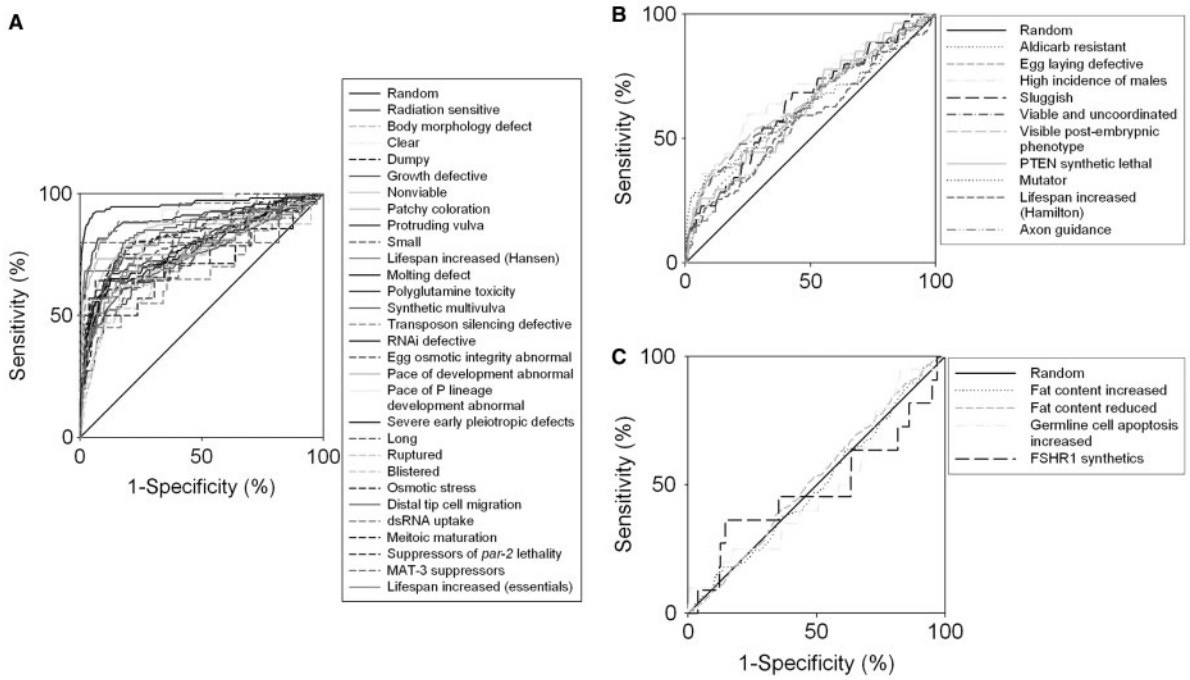

Figure 3: Using ROC curves to test the ability of a gene networks to predict loss-of-function phenotypes. The plots show ROC curves for the 'Wormnet' network of *C. elegans* using the data from 43 RNAi phenotypes. In this analysis, all genes in the genome are ranked by the sum of their interaction scores to a set of 'seed' genes that are known to result in each of the phenotypes when their expression is inhibited by RNAi. The curves show that Wormnet performs very well for predicting 29/43 phenotypes (**A**), better than random for a further 10/43 phenotypes (**B**) and the same as random for 4/43 phenotypes (**C**). This figure is adapted, with permission from reference [4].

been constructed to date, and we list examples for *C. elegans* in Table 2. Another important question is how best one can use these networks. The most trivial use of such a network is to look up a gene of interest and to see which genes are connected to it, what the evidence is for each of these connections, and then to use this information to guide future research. However, a more powerful way to use these networks is to search a network with groups of genes that are known to be important for a common process or phenotype [4, 47]. If you know a handful of genes that are important for a process, then it is much better to search a network using this group of 'seed' genes rather than using individual genes. For example, the web interface for Wormnet (http://www.functionalnet.org/wormnet/) allows the network to be searched using a list of query genes, returns a ROC curve that shows how well these query genes are clustered in the network (i.e. whether the network is predictive for these genes) and a list of all the genes that are predicted to interact with the query genes. The output also shows all the lines of evidence that are used to predict these interactions.

This use of gene networks to identify new candidate genes for a process of interest has been termed 'network-guided screening' [4]. Given knowledge of a few seed genes that are important for a process, a functional network can be used to predict a set of additional genes that can then be experimentally tested for their role in the process. To identify new candidate genes, every gene in the network can be ranked according to the sum of its connections to the known seed genes (just as in the ROC analysis described earlier in the article). Genes are then tested for their involvement in the process starting from the highest ranked genes and proceeding down the list. This approach dramatically reduces the number of genes that need to be tested. However, the approach should not entirely replace unbiased genome-wide screening, because genome-wide screens still offer the best method to identify entirely new pathways connected with a phenotype.

We provide two examples of how network-guided screening can work. The first example uses genes that have been identified in RNAi screens as increasing the lifespan of *C. elegans*. Three independent screens have been performed for this phenotype [50–52], allowing the validation rate of predictions made using the data from a single screen to be

Table 2: Integrated gene networks for *C. elegans*

Network	Reference	Description	Network availability
Co-expression network	[35]	A network of 22 163 evolutionary conserved co-expression relationships between ~3400 genes.	http://cmgm.stanford.edu/kimlab/multiplespecies
Early embryogenesis network	[13]	A network of interactions between 661 embryogenesis genes based on protein–protein interaction, phenotypic profiles and co-expression relationships.	Manuscript supplementary datafile
Zhong and Sternberg	[19]	An integrated network of 18 183 predicted genetic interactions between 2254 genes.	http://tenaya.caltech.edu:8000/predict
STRING	[8]	An integrated and regularly updated protein interaction and functional network for multiple species, including *C. elegans*.	http://string.embl.de/
Wormnet	[4]	An integrated functional network of 384 700 interactions between 16 113 genes.	http://www.functionalnet.org/wormnet/

estimated. For example, Hansen *et al*. [52] identified 29 genes that increase lifespan. If these genes are used to search the Wormnet network for additional genes, then of the top 50 novel genes connected to these genes, 10 (20%) are validated in one of the other two screens [4]. Of the top 200 most connected genes, 21 (10.5%) are validated in the other screens. Thus, screening the top 50 or 200 predictions made by Wormnet would have been between 100- and 4-fold more efficient in identifying longevity genes than performing genome-wide screens (depending upon the screen that is considered [4]).

The second example that we highlight concerns genes that function in the Retinoblastoma/Synthetic multivulva (Rb/SynMuv) pathway [53]. Previously, six genes were known that could suppress the phenotypic effects of mutations in this pathway [54, 55]. Searching the Wormnet network with these genes identified 62 and 142 connected genes in the high-confidence 'core' and lower confidence 'non-core' networks. On testing these genes it was found that 20% of the connected core and 5% of the connected non-core genes also acted as suppressors of the pathway. These numbers are 21- and 5-fold higher than the recovery rate from random screening [4]. Moreover, although Wormnet did not predict all the genes reported in a genome-wide screen as having this function [56], the genome-wide screen also did not report many of the genes identified using Wormnet. This is an important point—both network-guided screening and genome-wide screens have appreciable false-negative rates. Whereas for network-guided screening this is primarily because a network is incomplete (and also because not all genes that affect a phenotype are necessarily directly connected—the inactivation of two different biological processes may result in the same phenotypic change at the level of an organism), for genome-wide screens this is primarily because screening ~20 000 genes inevitably leads to an appreciable false-negative rate.

We recommend using all the data contained in a network—both the high- and low-confidence interactions—for network-guided screening. Provided that a network is probabilistic (i.e. that each interaction is weighted according to the confidence in it being correct), then it is advantageous to use both the weak and high-confidence data. This is because a novel gene may be connected by multiple weak lines of evidence to several of the seed genes used to search a network, resulting in a high overall ranking in the list of candidate genes. This does not necessarily increase the number of genes that need to be experimentally tested because genes that are only connected by one low-confidence interaction will still rank low on a list.

An additional advantage of using network-guided screening to identify new genes involved in a process is that the mechanistic interpretation of the 'hits' from a screen can be immediately aided by inspection of the evidence used that connects these genes to the seed genes in the network. For example, the presence of predicted protein interactions, co-regulation or orthologous interactions in other species can all greatly help to guide future mechanistic studies.

In a similar way, gene networks can also be used to help interpret the results of more traditional genome-wide screens, as they provide hypotheses for how the genes identified in a screen are functionally

connected to each other. Moreover, ROC curve analysis (as detailed above) can be used to determine how functionally 'coherent' the genes identified in a screen are. If interactions connect many of the genes identified in a screen, then they will receive a high AUC value, which suggests that given current knowledge these genes can be organized into a coherent 'pathway'. In contrast, if the genes are not very inter-connected in the network, then this implies either that we do not yet have sufficient data to mechanistically understand the process under investigation (implying the need for more experiments), or that the phenotype being investigated may be affected by mutations in many functionally unrelated genes.

CONCLUDING REMARKS

In this review, we have discussed how integrated functional gene networks have been built and used for model organisms such as yeast and *C. elegans*. In these species, genome-wide reverse genetic screens are relatively straightforward, which allows the predictive power of these networks to be extensively tested. Indeed, we hope that the *C. elegans* community will now start to use these networks to guide their research. However, the real power of network-guided screening will be its application to other organisms in which genome-wide screens are not readily possible. For example, the approach could be used in mice to identify new components of many pathways of interest.

Perhaps more promisingly the approach could also be applied to human biology. For most hereditary human diseases there are a limited number of genes that are known to be causally mutated in the disease. Using integrated gene networks it will be possible to use these known examples to predict new candidate genes that can then be tested for their association with a disease in population studies or in existing association-study data. In a typical genome-wide association study, hundreds of thousands of single nucleotide polymorphisms are tested for association with a disease phenotype. This multiple hypothesis testing greatly reduces the statistical power of these studies to identify real disease loci [57]. By predicting a limited set of functionally related candidate genes, network-guided screening may to some extent be able to overcome this limitation and provide more statistically powerful approaches for identifying human disease genes from association study data. Indeed, several recent papers have demonstrated how this approach may soon have a major impact on human genetics [58–60]. Most importantly, the work in model organisms has demonstrated that a single high-quality integrated network can be predictive for most of the different systems within an animal. Therefore, a single integrated network may be powerfully predictive for many different aspects of human biology and disease.

Key Points

- Functional gene networks connect two genes if they are likely to share the same function.
- These networks can be constructed by integrating many different types of experimental and genomic data together, according to the measured accuracy of each data set. The data sets used can come from many different organisms, and the final networks are probabilistic with each functional interaction associated with a confidence score.
- The predictive power of integrated gene networks can be tested using genome-wide reverse genetic data. *S. cerevisiae* and *C. elegans* are therefore currently the best systems for testing and validating integrated gene networks, and show that a single network can be predictive for all the cells and systems within a complex organism.
- Web interfaces to integrated gene networks make them accessible to all researchers online. This use of gene networks to find new genes important for a pathway of interest is termed 'network-guided screening'.

Acknowledgements

We thank Andrew Fraser and Edward Marcotte for leading us into the world of probabilistic networks and for all their ideas, help, guidance, support and collaboration over the years.

References

1. Bowers PM, Pellegrini M, Thompson MJ, *et al*. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 2004;**5**:R35.
2. Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* 2004;**32**:6414–24.
3. Lee I, Date SV, Adai AT, *et al*. A probabilistic functional network of yeast genes. *Science* 2004;**306**:1555–8.
4. Lee I, Lehner B, Crombie C, *et al*. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet* 2008;**40**:181–8.
5. 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.
6. Myers CL, Robson D, Wible A, *et al*. Discovery of biological networks from diverse functional genomic data. *Genome Biol* 2005;**6**:R114.

7. Troyanskaya OG, Dolinski K, Owen AB, *et al*. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA* 2003;**100**:8348–53.
8. von Mering C, Jensen LJ, Kuhn M, *et al*. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2007; **35**:D358–62.
9. Fraser AG, Marcotte EM. A probabilistic view of gene function. *Nat Genet* 2004;**36**:559–64.
10. Vidal M. A biological atlas of functional maps. *Cell* 2001; **104**:333–9.
11. Piano F, Gunsalus KC, Hill DE, *et al*. *C. elegans* network biology: a beginning. *WormBook* 2006:1–20.
12. Marcotte EM, Pellegrini M, Thompson MJ, *et al*. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999;**402**:83–6.
13. Gunsalus KC, Ge H, Schetter AJ, *et al*. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* 2005;**436**:861–5.
14. Lehner B, Fraser AG. A first-draft human protein-interaction map. *Genome Biol* 2004;**5**:R63.
15. Ramani AK, Bunescu RC, Mooney RJ, *et al*. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* 2005;**6**:R40.
16. von Mering C, Krause R, Snel B, *et al*. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;**417**:399–403.
17. Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;**302**:449–53.
18. Rhodes DR, Tomlins SA, Varambally S, *et al*. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* 2005;**23**:951–9.
19. Zhong W, Sternberg PW. Genome-wide prediction of *C. elegans* genetic interactions. *Science* 2006;**311**:1481–4.
20. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;**98**:4569–74.
21. Li S, Armstrong CM, Bertin N, *et al*. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004;**303**:540–3.
22. Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;**403**:623–7.
23. Walhout AJ, Sordella R, Lu X, *et al*. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 2000;**287**:116–22.
24. Gavin AC, Aloy P, Grandi P, *et al*. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006;**440**: 631–6.
25. Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002;**415**:180–3.
26. Krogan NJ, Cagney G, Yu H, *et al*. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006;**440**:637–43.
27. Byrne AB, Weirauch MT, Wong V, *et al*. A global analysis of genetic interactions in *Caenorhabditis elegans*. *J Biol* 2007;**6**:8.
28. Collins SR, Miller KM, Maas NL, *et al*. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007;**446**:806–10.
29. Davierwala AP, Haynes J, Li Z, *et al*. The synthetic genetic interaction spectrum of essential genes. *Nat Genet* 2005;**37**:1147–52.
30. Lehner B, Crombie C, Tischler J, *et al*. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat Genet* 2006;**38**:896–903.
31. Pan X, Ye P, Yuan DS, *et al*. A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 2006;**124**: 1069–81.
32. Schuldiner M, Collins SR, Thompson NJ, *et al*. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005; **123**:507–19.
33. Tong AH, Lesage G, Bader GD, *et al*. Global mapping of the yeast genetic interaction network. *Science* 2004;**303**: 808–13.
34. Marcotte EM, Pellegrini M, Ng HL, *et al*. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999;**285**:751–3.
35. Stuart JM, Segal E, Koller D, *et al*. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;**302**:249–55.
36. Huynen M, Snel B, Lathe W 3rd, *et al*. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000;**10**: 1204–10.
37. Pellegrini M, Marcotte EM, Thompson MJ, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999;**96**:4285–8.
38. Wolf YI, Rogozin IB, Kondrashov AS, *et al*. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 2001;**11**:356–72.
39. Enright AJ, Iliopoulos I, Kyrpides NC, *et al*. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999;**402**:86–90.
40. Yanai I, Derti A, DeLisi C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci USA* 2001;**98**:7940–5.
41. Matthews LR, Vaglio P, Reboul J, *et al*. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 2001;**11**:2120–6.
42. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**:D428–32.
43. Deplancke B, Mukhopadhyay A, Ao W, *et al*. A gene-centered *C. elegans* protein-DNA interaction network. *Cell* 2006;**125**:1193–205.
44. Harbison CT, Gordon DB, Lee TI, *et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;**431**: 99–104.
45. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.

46. Wong SL, Zhang LV, Tong AH, *et al*. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* 2004;**101**:15682–7.

47. McGary KL, Lee I, Marcotte EM. Broad network-based predictability of *S. cerevisiae* gene loss-of-function phenotypes. *Genome Biol* 2007;**8**:R258.

48. Lehner B, Fraser AG, Sanderson CM. Technique review: how to use RNA interference. *Brief Funct Genomic Proteomic* 2004;**3**:68–83.

49. Cerón J, Cabello J, Monje JM, *et al*. *Applications of RNAi in C. elegans Research. RNAi Interference Research Progress*. New York: Nova Publishers, 2008.

50. Curran SP, Ruvkun G. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet* 2007;**3**:e56.

51. Hamilton B, Dong Y, Shindo M, *et al*. A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev* 2005;**19**: 1544–55.

52. Hansen M, Hsu AL, Dillin A, *et al*. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS Genet* 2005;**1**:119–28.

53. Ferguson EL, Horvitz HR. The multivulva phenotype of certain *Caenorhabditis elegans* mutants results from defects in two functionally redundant pathways. *Genetics* 1989;**123**: 109–21.

54. Lehner B, Calixto A, Crombie C, *et al*. Loss of LIN-35, the *Caenorhabditis elegans* of the tumor suppressor p105Rb, results in enhanced RNA interference. *Genome Biol* 2006;**7**:R4.

55. Wang D, Kennedy S, Conte D, Jr, *et al*. Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. *Nature* 2005;**436**:593–7.

56. Cui M, Kim EB, Han M. Diverse chromatin remodeling genes antagonize the Rb-involved SynMuv pathways in *C. elegans*. *PLoS Genet* 2006;**2**:e74.

57. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;**7**:781–91.

58. Franke L, Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**:1011–25.

59. Lage K, Karlberg EO, Storling ZM, *et al*. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;**25**:309–16.

60. Pujana MA, Han JD, Starita LM, *et al*. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* 2007;**39**:1338–49.

61. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.