# 8 Inferring Genetic Architecture from Systems Genetics Studies

Xiaoyun Sun[3], Stephanie Mohr[1], Arunachalam Vinayagam[1],Pengyu Hong[3] and Norbert Perrimon[1,2,4]

1 Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.

2 Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.

3 Department of Computer Science, Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02454, USA.

4 Corresponding author: perrimon@receptor.med.harvard.edu

In recent years many efforts have been invested in comprehensively evaluating the behavior and relationships of all genes/proteins in a particular biological system and at a particular state. Here, we review how genome-wide RNAi screens together with mass spectrometry can be integrated to generate high confidence functional interactome networks. Next we review the mathematical modeling methods available today that allow the computational reconstruction of such networks. Network modeling will play an important role in generating hypotheses, driving further experimentation and thus novel insights into network structure and behavior.

## 8.1 Introduction

Most biologists study a specific biological problem by investigating the activities of a limited number of genes or proteins involved in a particular biological process. This traditional approach is critical and has proven to be extremely successful to reveal the detailed molecular functions of individual genes and proteins. For example, genetic studies of embryonic patterning in Drosophila identified about 40 genes with striking segmentation defects that fell into distinct phenotypic classes: gap genes, pair rule genes, segment polarity genes, and homeotic genes (Nusslein-Volhard & Wieschaus 1980). Detailed analyses of the mutant phenotypes and functions of even this relatively small set of genes led to a comprehensive molecular framework of the process of embryonic patterning (St Johnston & Nusslein-Volhard 1992). Reductionist approaches, however, are not sufficient for generating the big picture of how a biological system, including multiple levels of many different gene products and the interactions among them, works at different physiological states or developmental stages (Friedman & Perrimon 2007). Thus, as our knowledge of individual genes and proteins accumulates, there is a need to comprehensively evaluate the behavior and relationships of all genes/proteins in a particular biological system and at a particular state. In recent years, progress has been made in multicellular organisms towards this goal mostly in tissue culture, a platform that allows a sufficient amount of homogeneous material to be easily obtained.

Profiling the parameters involved in various biological processes at genome scale has become a promising strategy to address such a Systems Biology problem. This approach is now possible due to advances in RNA interference (RNAi), whereby the functions of all annotated genes in a genome can be systematically interrogated (Mohr et al. 2010). Furthermore, major technical advances in proteomics, transcriptomics, and cellular imaging now provide sophisticated means to measure biological parameters quantitatively and at high-throughput scale. Altogether, these approaches allow the generation of phenotypic signatures for all genes expressed in a cell of interest that describe their roles in the biological process under scrutiny. The goal of applying these methods is not only to provide functional information on the activities of many genes/proteins, but also to enable the construction of networks that faithfully reflect the dynamics of biological activities in a particular system. This approach is challenging as both biological and technical noise can affect the quality of the data sets generated, and requires in particular robust cellular assays, careful consideration of the reproducibility of the data generated, integration of orthogonal data sets, and rigorous computational analyses.

Three types of experimental data sets are most frequently integrated in network construction: transcriptomics, proteomics and interactomics. Transcriptomics provides information about both the presence/absence and relative abundance of RNA transcripts, thereby indicating the active components within the cell. Transcriptome data measured by genome-wide microarray or RNA-seq (transcriptome profiling that uses deep-sequencing technologies) is widely used for network construction, as RNA molecules are easily accessible in comparison to proteins and metabolites. Proteomics describes the entire population of expressed proteins in a cell or tissue. It aims to identify and quantify the cellular levels of genome-wide protein expression in a specific biological system. Interactomics include protein-DNA, protein-RNA and protein-protein interactions (PPIs). Protein-DNA interactions mainly occur between transcription factors and their target DNA, whereas protein-RNA interactions depict potential regulatory roles of specific proteins to target RNAs. PPIs define the fundamental genetic regulatory network of the cell. They are extremely valuable for network construction, as with this approach the relationships among interacting proteins are clearly established, in contrast to the often indirect and sometimes complicated regulation of components within genetic networks. Finally, in addition to transcriptomics, proteomics and interactomics, the analysis of phenotypic signatures, based on cellular features extracted from image analyses, has emerged as a powerful method that provides rich phenotypic information on dynamic and more complex cellular processes, such as nuclear translocation, cytokinesis and cell migration (Perlman et al. 2004, Bakal et al. 2007).

Here, we review how some of these methods can be applied today to Drosophila cells to gain insights on the organization of biological networks. Specifically, we describe how genome-wide RNAi screens are used to identify gene activities in the cell that affect the output of a network, then we describe how PPIs, generated from Mass Spectrometry, can be easily integrated with RNAi data sets to generate a high confidence functional interactome. Finally, we review the mathematical modeling methods available which, when applied to the integrated data sets generated from RNAi and PPI, allow the computational reconstruction of the network - the goal of which being to generate a number of hypotheses ultimately driving further experimentations leading to novel insights.

## 8.2 Identification of network components by RNAi

In *Drosophila* cells, RNAi knockdown is easily achieved using in vitro synthesized long dsRNAs (typically  150 to 500 bp), and readily adaptable to screening of cultured or primary cells in miniaturized platforms (e.g. 384-well plates) (Boutros et al. 2004). Thus, *Drosophila* cell-based RNAi screening can be done in high-throughput mode, providing a platform for genome-scale functional analysis of cellular processes (DasGupta & Gonsalves 2008, Bakal & Perrimon 2010, Falschlehner et al. 2010, Mohr et al. 2010). Information about reagents and results from *Drosophila* cell-based RNAi screens is available at a number of databases, including FLIGHT (http://flight.icr.ac.uk/) (Sims et al. 2006), GenomeRNAi (http://genomernai.de/GenomeRNAi/) (Gilsdorf et al. 2010) and FlyRNAi, the database of the *Drosophila* RNAi Screening Center (http://www.flyrnai.org) (Flockhart et al. 2006). To date, large number of screens have been performed in *Drosophila* cells, yielding insights into a number of biological processes and systems (Mohr et al. 2010). Researchers often screen grouped sub-sets of genes, e.g. all genes encoding kinases or genes identified using another high-throughput method or bioinformatics analysis. However, full-genome screening remains the most unbiased and comprehensive approach. An important aspect of RNAi screening distinguishing it from some other high-throughput methods is that RNAi results not only implicate genes in a given pathway but can also indicate the direction of action (i.e. a positive or negative regulator in a given pathway).

A wide variety of high-throughput screening methodologies, instruments and assays are available for RNAi screening ((Shumate & Hoffman 2009, Mohr et al. 2010); Figure 1). Among the most straightforward to perform and analyze are total-well luciferase or fluorescence readouts, which are collected using a luminometer or fluorimeter (plate-reader). These outputs are typically expressed as a ratio (e.g. of values obtained with a transcriptional reporter versus a ubiquitously expressed control). From these numerical outputs, positive hits are typically identified after calculating Z-scores and choosing an appropriate Z-score cut-off value. Researchers often rely on prior knowledge of components of a process or system in order to select an appropriate Z-score value. This is often done empirically (i.e. I know gene X is involved and reagents targeting X gave a Z-score of n in the screen; therefore, I will use n as my cut-off). However, it can also be done more systematically, such as using RNAiCut, which is based on the assumption that gene products corresponding to true positives are more interconnected to one another at the level of PPIs (Kaplow et al. 2009). A number of other types of assays, supported by specialized plate-readers or laser-scanning cytometers, are similar in that all cells in the well are measured and the data is in the form of one or more numerical outputs that can be analyzed based on Z-scores. Among these, the in cell Western approach, in which immunofluorescence labelling of a phospho-protein or protein is compared to a total protein control, has proved particularly useful for interrogation of signal transduction in *Drosophila* cells (Friedman & Perrimon 2006, Friedman & Perrimon 2007, Kockel et al. 2010, Friedman et al. 2011).

Although relatively simple outputs continue to be informative for screening, researchers are increasingly turning to high-content image-based screens in order to obtain high-quality results relevant to complex phenotypes (Bakal 2011). Instruments developed for acquisition of high-content screen image data include automated epifluorescence, fluorescence confocal and laser-scanning microscopes (Shumate & Hoffman 2009, Zanella et al. 2010). Most of these instruments image a sub-region of the well and multiple images per well must be acquired in order to image enough cells for statistically meaningful results. Through the use of one or more fluorescent dye or antibody, as well as the introduction of fluorescence protein-tagged fusion proteins or reporters, several different readouts can be simultaneously collected and measured. Even a single image-based readout such as the DNA dye 4',6-diamidino-2-phenylindole (DAPI) can be used to count cells, define the nucleus (e.g. as a reference for detecting nuclear vs. cytoplasmic localization), measure nuclear area, monitor cell cycle stages, and more. As screen-imaging instruments facilitate collection of data for several different fluorescent tags, the number of features that can be evaluated singly or in relationship to one another can be very large. Thus, high-content screening facilitates detection of complex cellular and sub-cellular phenotypes, such as changes in the sub-cellular distribution of a protein, or in the size, shape or number of cells or organelles (see for example (Bakal et al. 2007)). Importantly, analysis of multiple parameters can be used to improve the quality of RNAi screen results. Assessment of the Z factor for analyzed images during assay development, for example, can be used as a measure of robustness prior to the screen, guiding optimization of the screen assay (Kummel et al. 2010). Moreover, following image data acquisition, some parameters prove more informative than others in identifying on-target and relevant cellular responses (Collinet et al. 2010, Kummel et al. 2010). Development in the area of screen image analysis is growing rapidly. Various academic and commercial groups have developed software tools useful for analysis of high-content screen datasets, including machine-learning approaches (reviewed in (Ljosa & Carpenter 2009, Niederlein et al. 2009) (Ljosa and Carpenter 2009; Niederlein, Meyenhofer et al. 2009)). Following analysis,

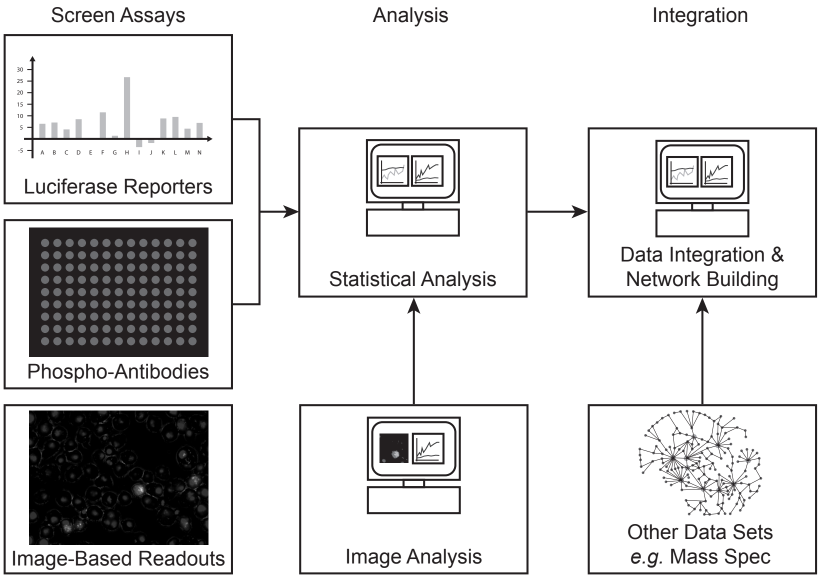


**Figure 8.1** Cell-based assays used to describe cellular phenotypes. Several types of assays and corresponding instruments are available for RNAi screening and other high-throughput *Drosophila* cell-based methods. Luciferase levels provide a relatively simple way to monitor overall induction or suppression of a transcriptional reporter. Methods such as the in cell Western approach allow for monitoring of endogenous proteins. With this approach, immunofluorescence levels of a phospho-protein or protein are compared with levels detected using non-phospho-specific antibody or total protein dye. High-content imaging allows for detection of sub-cellular protein distributions, organelles and other features. For high-content imaging, epifluorescence or confocal microscopy is used to detect one or several cellular and sub-cellular readouts, followed by single- or multi-parametric image analysis of simple or complex phenotypes. The complexity of image data requires specialized analysis. In all cases, phenotypes are reduced through analysis to numerical values such as Z-scores. These results can then be combined with results from other high-throughput assays, e.g. mass spectrometry, or with information from the literature in order to build high-confidence gene networks.

the screen output is reduced to one or more numerical values, which can then be evaluated using Z-scores or another approach.

Despite the power of the RNAi approach, a variety of caveats apply that are relevant to the analysis, interpretation and integration of RNAi screen results in the context of system-wide analyses. Perhaps the most common problems are systematic errors or bias in the assay; stochastic effects or noise inherent in high-throughput data sets; and reagent-specific off-target effects (Falschlehner et al. 2010, Mohr et al. 2010, Booker et al. 2011, Seinen et al. 2011). Most systematic errors are easily addressed prior to primary data acquisition (e.g.

through correction of instrument-derived dispensing errors) but others persist and can affect interpretation of results. For example, some cell-based assays show bias, such as favouring identification of hits in one direction or the other relative to controlse.g. favouring identification of positive-acting or negative-acting factors in a pathway (DasGupta et al. 2007). When such a bias exists and cannot be fully addressed through assay optimization, robust interrogation of a given network might require screening with more than one assay, i.e. by performing related assays with opposite biases. In general, screening of multiple time-points or conditions, as well as screens that combine more than one dsRNA reagent to look at additive or synergistic effects of double knockdown, can help reduce false negative discovery. For example, genes whose knockdown results in weak phenotypes not detected above noise in a single knockdown screen might be picked up as significant positives when combined in a double-knockdown screen (Bakal et al. 2008).

Regarding the identification of false positives in RNAi screens, a number of approaches are available (Figure 8.2). The most rigorous approach to validation of screen hits is a rescue test, in which a construct that confers gene activity but can evade the RNAi reagent is tested for the ability to reverse the observed phenotype. For many screens, initial positive hits are re-tested with two or more unique reagents per gene in order to filter out potential false positive results. An additional filter that can be applied is to remove initial positive hits for which there is no evidence of gene expression in the cell line that was tested, with the underlying assumption that reagents targeting genes for which there is no evidence of expression are more likely to be exerting their effects through off-targets (Booker et al. 2011). Informatics-based analysis of reagent quality (e.g. number of predicted off-targets) can also be taken into account in assessing primary screen data. However, the experimental approaches are impractical to apply at large scale, and the systematic approaches can limit but not eliminate false discovery. As a result, systematic and robust detection of false positive and negative results is simply not practical to do for all genes tested in a large-scale RNAi study. Thus, in general, a researchers curated list of positive results from an RNAi screen is likely to be based upon a combination of statistical analysis, experimental verification, and/or prior knowledge of the process or pathway under study, and genome-wide information is usually only available in the form of analyzed but unverified primary screen data (e.g. Z scores for all primary hits). In addition to these methods overlapping orthogonal data sets with RNAi results, as describe below, provide a powerful filter to identify high confidence network components.

## 8.3 Identification of network components using Proteomics

PPIs play critical roles in many cellular processes, such as signaling cascades and regulatory complex formation. In addition, information acquired from PPI

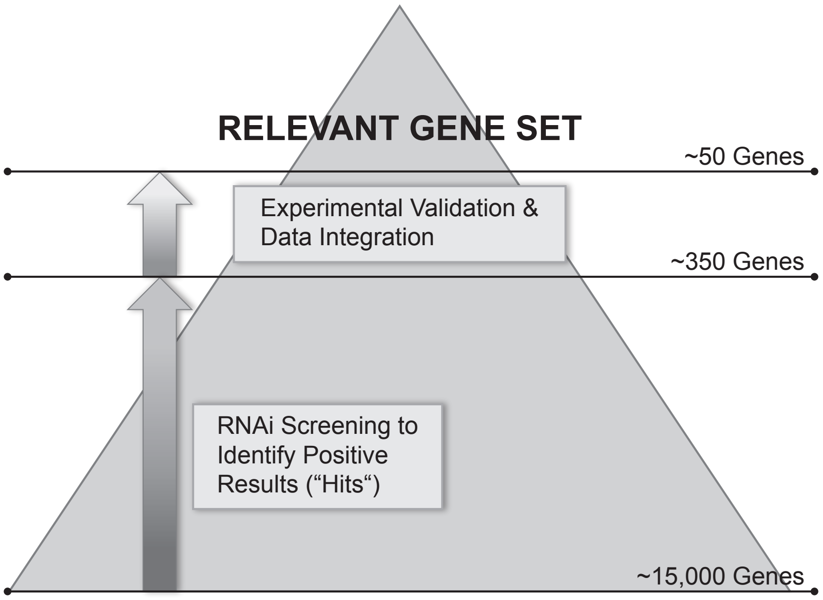

RELEVANT GENE SET
~50 Genes
Experimental Validation & Data Integration
~350 Genes
RNAi Screening to Identify Positive Results (“Hits“)
~15,000 Genes

**Figure 8.2** Validation and integration of RNAi screen data can help identify high-confidence gene networks. Full-genome screening allows researchers to identify screen hits (positive hits), reducing the candidate gene list to a more manageable size. Subsequent to the screen, experimental validation, as well as integration and filtering with other data sets, can be used to identify a high-confidence set of genes likely to be involved in a given process or pathway. Commonly used methods for experimental validation include testing with more than one unique dsRNA reagent per gene, testing for concordance between quantitative reverse transcriptase PCR (qPCR) analysis of mRNA knockdown and observed phenotypes, and RNAi rescue tests. Commonly used computational approaches include comparison of the set of screen hits to evidence for gene expression (with the assumption that reagents targeting genes known to be expressed are more likely to be exerting on-target effects), and integration with the results of other high-throughput studies, such as mass spec analyses, or information culled from the published literature.

data is definitive (i.e., proteins A and B interact with each other) and can contain quantitative features (i.e., the strength of interactions), making PPI data a core resource for network construction. Hence identifying all functional PPIs is not only important for understanding the structure and function of biological systems, but also for the construction of reliable networks.

Although classic biochemical experiments based on co-immunoprecipitation can readily identify interactions between specific proteins, they lack the ability to explore interactions at whole proteome scale. In recent years, several techniques measuring proteome-wide PPIs have been developed. Two methods in particular facilitate high-throughput studies, the yeast two-hybrid approach (Fields 2005) and tandem affinity purification coupled with mass-spectrometry (TAP/MS) (Gavin et al. 2006, Krogan et al. 2006, Jeronimo et al. 2007). The yeast two-hybrid approach is a useful approach for high-throughput identification of putative interaction partners, but it can be prone to false-positive identification and interactions are detected in a heterologous context. Thus, TAP/MS analysis has been increasingly used to identify novel and large-scale PPI under physiologically relevant conditions (Gavin et al. 2006, Krogan et al. 2006, Jeronimo et al. 2007).

TAP and MS are two essential components of the TAP/MS technique. TAP efficiently isolates native protein complexes from cells for proteomics analysis. It is followed by MS analysis, a powerful analytical technique used to determine the molecular structures of peptides. The advantage of MS is that it identifies multi-subunit protein complexes isolated from the cell lysate with extremely high sensitivity and accuracy. The TAP/MS approach has been used successfully to characterize protein complexes from various cells and multi-cellular organisms. In addition, this technique can be combined with quantitative proteomics approaches to better understand the dynamics of protein-complex assembly. As will be discussed below, TAP/MS can also be integrated with RNAi data, so that high confident and even dynamic networks can be reconstructed.

To study pathway-specific interactions, special cell lines need to be generated first, with each cell line stably expressing a TAP-tagged version of a starting protein of interest (the bait protein), such as a major signaling component. The reason for tagging the components is to facilitate isolation of those components later. Along with the special cell lines, a negative control cell line (i.e., not expressing any TAP-tagged proteins) is recommended for subtracting nonspecific interactors. Both types of cells are treated with specific conditions to produce proteins lysates. The lysates are incubated with affinity purification beads, where the TAP-tagged protein is pulled down via its tag, together with associated proteins (the prey proteins) and other proteins retained through non-specific binding. The protein samples collected are then broken down into peptides with proteases and analyzed by MS, where a list of peptide sequences from each sample is reported as the results. A necessary data pre-processing step is to identify the source proteins of the peptide sequences and calculate the number of peptides for each prey protein identified in each sample. To increase confidence in PPI identification, multiple replicates are recommended for each cell line and condition.

Early TAP/MS analytic methods identify PPIs by binary mode (i.e., indicating the presence or absence of a specific protein) (Zhu et al. 2007). Newer methods take into account quantitative information such as the label-free quantitative spectral count (SC), which is the number of peptides detected in MS. The challenge for TAP/MS data analysis is to minimize false-positive interactions and increase the sensitivity to identify true interactions. Currently, there are three popular computational tools for TAP/MS data analysis: NSAF (Normalized Spectral Abundance Factor) (Sardiu et al. 2008); CompPASS (Comparative Proteomic Analysis Software Suite) (Sowa et al. 2009); and SAINT (Significance Analysis of Interactome) (Choi et al. 2011).

NSAF estimates the relative abundance of proteins based on the total number of peptides (i.e. SC) identified in the sample. In general, larger proteins are expected to generate more peptides and hence a larger SC than smaller proteins. To account for the variation of protein size, the SC for each protein is divided by the protein length, which is defined as the spectral abundance factor (SAF). Individual SAF also needs to be normalized by the sum of all SAFs for proteins in the sample to accurately account for run-to-run variation (Eq. 8.1).

$$NSAF(i) = \frac{\left(\frac{SC_i}{L_i}\right)}{\sum_{i=1}^{N}\left(\frac{SC_i}{L_i}\right)} \tag{8.1}$$

The NSAF for a protein i is the SC of a protein divided by the proteins length (L), divided by the sum of SC/L for all N proteins in the experiment. NSAF is simple, easy to compute and has been demonstrated to be effective in detecting significant PPI. However, NSAF is an empirical transformation of SCs, which does not incorporate any information from negative controls. Moreover, it does not add weight to interactions that are detected in all biological replicates (most likely true interactions) and does not penalize interactors detected in all purifications (e.g. sticky proteins that interact with all bait proteins). Thus, although NSAF is useful to some extent, it clearly needs further improvement.

The CompPASS method computes PPI scores by adjusting observed SCs relative to the reproducibility of detection across biological replicates, as well as the frequency of observing the prey protein in purifications with different baits. The first step in CompPASS is the generation of a Stats Table (Table 8.1). In the table, each row is the unique protein identified from the TAP/MS experiments (interactor) and each column is the bait protein used in those experiments. Each element of the table is the SC of an interacting protein from the particular baits TAP/MS experiment. After the stats table is created from all experiment runs in the project, CompPASS calculates a mean value of the SC (M) for each interactor, then calculates a Z-score and D-score for each interaction pair.

**Table 8.1** Stats Table in ComPASS analysis

| | Bait1 | Bait2 | Bait3 | Bait4 | Bait k | Mean |
|---|---|---|---|---|---|---|
| Interactor 1 | $X_{1,1}$ | $X_{1,2}$ | $X_{1,3}$ | $X_{1,1}$ | $X_{1,k}$ | $M_1$ |
| Interactor 2 | $X_{2,1}$ | $X_{2,2}$ | $X_{2,3}$ | $X_{2,4}$ | $X_{2,k}$ | $M_2$ |
| Interactor 3 | $X_{3,1}$ | $X_{3,2}$ | $X_{3,3}$ | $X_{3,4}$ | $X_{3,k}$ | $M_3$ |
| Interactor n | $X_{n,1}$ | $X_{n,2}$ | $X_{n,3}$ | $X_{n,4}$ | $X_{n,k}$ | $M_n$ |

The first score is the Z-score, which is specific for a particular interaction; the mean is subtracted from the SC, and is divided by the standard deviation ( Eq. 8.2 and 8.3). X is the SC, i is the bait number, j is the index of interactor,

**Table 8.2** Advantages and disadvantages of network models

| | Information theory model | Boolean Network | Differential Equations Model | Bayesian Network | Dynamic Bayesian Network |
|---|---|---|---|---|---|
| Simplicity | Yes | Yes | No | No | No |
| Low computational cost | Yes | Yes | No | No | No |
| Multiple genes participate in one function | No | Yes | Yes | Yes | Yes |
| Directed/ Undirected | Undirected | Directed | Directed | Directed | Directed |
| Large dataset needed | No | No | Yes | Yes | Yes |
| Deterministic/ Stochastic | Deterministic | Deterministic | Deterministic | Stochastic | Stochastic |
| Handle incomplete data | No | No | No | Yes | Yes |
| Handle feedback loops | No | No | No | No | Yes |

$n$ is the total number of interactors, $k$ is the total number of baits, $M$ is the mean of the SC and $\sigma$ is the standard deviation of the SC for each interactor. Although the Z-score can identify interactors for which the SC is significantly different from the mean, it fails to discriminate two interactors with dramatically different SCs if the experiment has only one replicate. For example, if in a single experiment, the SCs for A and B SC are 2 and 20, respectively, then the two proteins will have the same Z-score, as the mean and standard deviation are the same for a single data point.

$$Z_{ij} = \frac{X_{ij} - M_i}{\sigma_i}, \text{ where } M_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij}. \tag{8.2}$$

The second is the D-score (Eq. 8.3), which takes into account both the reproducibility of detection across biological replicates and the frequency of observing prey protein in purifications of different baits. The variables are the same as for Eq.s 8.2 and 8.3. Here, $f$ is a term which is 0 or 1 depending on whether or not the interactor was found a given particular bait. $\sum f$ is the summation across all baits. $k/\sum f$ is the frequency of this particular interactor across all baits. P is the number of replicate runs in which the interaction is present. The reproducibility term allows for better discrimination between a likely false positive (i.e. an interactor found in a one run but not in any of the other multiple replicates) and a likely true positive (i.e. an interactor with a low SC yet found in all replicates).

$$D_{i,j} = \sqrt{X_{ij}\left(\frac{k}{\sum_{i=1}^{k} f_{ij}}\right)^{p}} \quad \text{where} f_{ij} = \begin{cases} 1 & \text{if } X_{ij} > 0 \\ 0 & \text{else} \end{cases} \tag{8.3}$$

CompPASS is easy to compute and takes into consideration two important factors: reproducibility and the frequency at which each interactor is detected in multiple replicates. It uses a different approach to distinguish the background and real interactors rather than directly utilizing the negative control datasets. Further, CompPASS takes a maximum of 2 replicates which might not be enough for some experiments with large variance in their biological replicates.

The SAINT approach assigns a confidence score to a PPI by converting the normalized SC into the probability of a true interaction between the two proteins. The parameters for true and false distributions, $P(X_{ij}|\text{true})$ and $P(X_{ij}|\text{false})$, and the prior probability of interactions in the dataset, $P(\text{true})$ and $P(\text{false})$, are inferred from the normalized SCs for all interactions that involve prey $i$ and the bait $j$. The posterior probability of a true interaction, $P(\text{true}|X_{ij})$, can be calculated from parameters using Bayes rule (Eq. 8.4).

$$P(\text{true} \mid X_{ij}) = \frac{P(X_{ij}|\text{true})P(\text{true})}{P(X_{ij}|\text{true})P(\text{true}) + P(X_{ij}|\text{false})P(\text{false})} \tag{8.4}$$

SAINT modeling can be performed with or without negative control data. When negative controls are not available, the distribution of false interactions can be estimated in reference to the quantitative information for the same interactor across purifications of all other baits in the dataset. When TAP/MS data contains negative controls, SAINT estimates the SC distribution for false interactions directly from the negative controls. The incorporation of negative control data improves the robustness of modeling, especially for small datasets.

The SAINT model is based on label-free quantification using the SC. It constructs separate distributions for true and false interactions to derive the probability of a bona fide PPI. The probability model can also be used to estimate the false discovery rate (FDR), and can be extended to model other types of quantitative parameters such as peptide ion intensity. However, SAINT specifically excludes proteins with 1-2 SCs. The necessity of this arbitrary step is questionable. Moreover, the complicated reference procedure in SAINT demands high computational costs.

Overall these approaches can effectively analyze TAP/MS datasets, but they also have room for improvement. For example, NSAF and CompPASS compute scores based on the transformation of SC, and SAINT demands high computational costs largely due to its complicated reference procedure. New algorithms should be investigated in the future that eliminate false positives more effectively and that require lower computational costs.

## 8.4 Integration of RNAi and Proteomic data sets

RNAi provides information about which genes affect the activity of a network. However, many gene activities can affect the activity of a network indirectly (e.g., general maintenance activities such as those related to overall protein translation or stability). In addition, as discussed above, although methods are available to ensure that RNAi effects are on target, such as genomic rescue, the most rigorous validation approaches can be tedious and are not commonly used large-scale for validation of RNAi results. Proteomic data sets, which frequently reflect the interactions among different proteins at a genome-wide scale, can help address both of these issues, as the integration of RNAi and proteomic data sets can facilitate validation, leading relatively quickly to a high confidence functional interactome.

Different data sets are usually ranked using different types of scores (e.g., Z-score is used to evaluate the result of RNAi screens; a probability value or p-value is generated from TAP/MS data sets by various analytic methods). Thus, the first challenge in integrating RNAi and PPI results is to combine different data sets with different scoring functions. One common approach is to choose appropriate cut-off value (threshold) for each data set and integrate them using the voting system (Zhong & Sternberg 2007), in which a simple statistical model is used to integrate multiple data sets in the absence of data training. Basically, with this approach, gene/protein that appears in each data set gets one vote, and the total number of votes are computed and used to determine either inclusion or exclusion from the network. When the threshold vote number equals the total number of data sets (i.e., scored in all data sets), the system becomes a filtering model and only the intersection of all data sets is selected. When the threshold is set to one, the system selects the union of all data sets. The threshold directly affects both the false positive and false negative rates in the final data integration results, and should be set according to different analytic purposes. A small value of the threshold will give rise to a relative complete network, but more errors might be associated. On the other hand, high threshold value can generate a high confident network, but it might also eliminate some useful information.

Because of the simplicity of the voting system, and because there is no requirement for a training data set, it has been extensively used in a variety of investigations (Walhout et al. 2002, Gunsalus et al. 2005). For example, in a recent study of the canonical receptor tyrosine kinase (RTK)/RAS/extracellular signal-regulated kinase (ERK) pathway in *Drosophila* (Friedman et al. 2011), a comprehensive network was integrated by combining unbiased ERK activation genome wide RNAi screens with TAP/MS network structural data. In this study, RNAi screen results were filtered using interactors identified in the PPI network in order to achieve significant enrichment of pathway regulators. The results showed that about 50% the filtered PPI network scored in the RNAi screens.

The integration of multiple data sources improves the specificity and reliability of individual high-throughput data sets. It can also be an effective approach to reduce the level of false negative discovery (i.e. protein complexes identified in the network reconstruction can guide experimental validations of some interactions not scored in the original TAP/MS data sets). Furthermore, by combining RNAi and TAP/MS data sets with time-course measurements, aspects of the dynamic regulation of the network can be revealed.

## 8.5 Network modelling: The next step

Following the construction of a network involved in a particular biological process (e.g., the *Drosophila* RTK/ERK network; (Friedman et al. 2011)), involves network reconstruction using mathematical modeling. Network reconstruction aims to build a mathematical model through a learning algorithm, so that the output of the model fits with provided biological data, and the relationships of the network components (genes/proteins) are clearly defined. Essential in this network reconstruction process is a solid computational analysis, which involves data preparation, network architecture selection, and structure and parameter learning. Data preparation is the fundamental step and largely determines the quality of the analysis outcome. Appropriate network model selection depends on both the available data type and the aims of the computational analysis. The final network can be built through a repetitive structure and parameter learning and refining processes (illustrated in Figure 8.5). A good network not only depicts the detailed regulation of its components but also provides high-confidence and promising directions for future experimental design.

### 8.5.1 Data Preparation

Good data preparation is key to network reconstruction and a balancing act. On one hand, in order to minimize experimental efforts and costs, the number of experiments conducted should be minimal; on the other hand, accurate reconstruction of biological network demands a considerable quantity of reliable data. In determining the amount of data required for network inference, the complexity of the system and the quality of the network are integral and related factors (Hecker et al. 2009). Generally, the quality of a network largely depends on the quality of the given data. Large variation and high levels of measurement noise in the experiment data will impair the quality of constructed network. Thus, it is important to carry out multiple replicates to minimize the effects of variation and noise. Incorporation of a larger number of parameters allows a network to more accurately represent the complexity of a system; however, this requires collection of a larger amount of experimental data and also adds to the total computational time.

Data pre-processing is an important step in data preparation. It directly affects both the performance of the network inference algorithms and the inference results. Methods for data preprocessing need to be applied selectively according

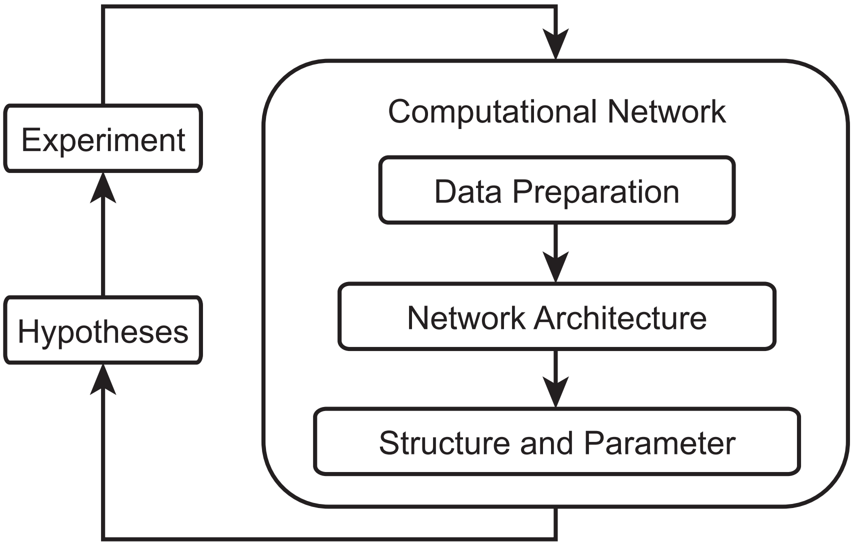


**Figure 8.3** Biological experimenting and computational network reconstructing cycle. Hypothesis-driven biological experiments are analyzed by computational approaches, aiming to reconstruct underlying network. This involves Data preparation, Network Architecture Selection, Structure and Parameter Learning steps. The inferred network improves our understanding of biological systems and further aids the guidance for future experimental designs. A new round of experiments enables further improvement of subsequent network construction.

to different data types, experimental designs, and network inference methods. For instance, certain methods only allow for input of binary numbers, so measured expression levels have to be converted into two discrete expression values. Other methods require time-series data, so the appropriate interpolation of experiment data at different time points has to be conducted during data preprocessing.

In general, to construct a reliable network while also limiting network complexity and computation time, the following strategies should be considered in data preparation (Hecker et al. 2009). First, the amount of data should be increased either by increasing the number of measurements or through data integration. Second, the number of network components should be reduced by grouping together genes/proteins with similar functions. Third, the number of network parameters should be restricted so that the dimensionality of the network search space can be reduced. And finally, specific prior knowledge from various sources should be incorporated to reduce the number of parameters.

### 8.5.2 Network Model Selection

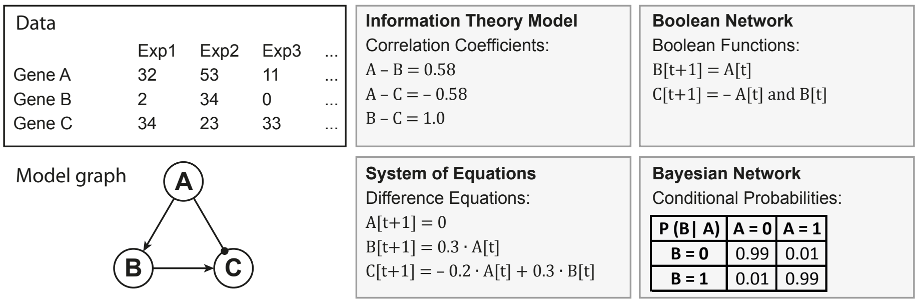
To reconstruct a network, it is important to start with an appropriate type of network model. Network model adopts the mathematical function to depict the general behavior of the network components. Most network model can be represented by a graph containing both nodes and edges. The nodes represent network components (e.g. genes, proteins or protein complexes), and edges between nodes represent the interactions between network components. Edges are either directed (indicating the directionality of the interaction, for example, if we have $A \to B$, node $B$ is regulated by node $A$) or undirected (indicating presence/absence of the interaction, for example, if we have $A$–$B$, nodes $A$ and $B$ interact). Once the network model is defined, details of the model will be learned from the experimental data: the network structure illustrates the interactions among all the components in the system and the model parameters characterize different aspects of the interactions, e.g. their types/strength.

Several network models have been proposed over the past few years. These models make distinct assumptions about the underlying molecular mechanisms with varying degrees of simplification. In these network models, the activity of a component can be represented by Boolean (0 or 1), discrete (e.g. 1, 2, 3), or continuous (real) values; the type of relationships between the variables (A and B) can be directed ($A \to B$) or undirected ($A$—$B$), linear (e.g. $A = \alpha_1 B + \alpha_0$) or non-linear (e.g. $A = B^2$). The type of model can be deterministic or stochastic, static or dynamic. A deterministic model always predicts the same outcome when the initial conditions are the same, whereas a stochastic model characterizes the probability distribution of possible outcomes. Dynamic models generally define a parametric model of interactions and try to estimate the parameters from different time points (e.g. time course gene expression data). Static models characterize causal interactions that are consistent across the measurement (van Someren et al. 2002).

Currently there are five distinct and widely used network models (Figures 8.5.2 and 8.5.2): Information Theory model, Boolean Network, Differential Equation Model, Bayesian Network (BN) and Dynamic Bayesian Network (DBN). The strengths and weaknesses of these network models will be addressed below (summary in Table 8.2).

The Information Theory Model is one of the simplest network models 'citeStuart2003. It represents the regulatory system with an undirected graph, in which nodes represent components of the system and edges are interactions between components. Simplicity and low computational costs are the major advantages of information theory models. It has been widely applied to study global properties of large-scale regulatory systems. However, a drawback of this model is that it is static and cannot adequately account for complex regulation involving multiple gene/protein components.

A Boolean network is a discrete dynamical network (Kauffman 1969). It can be represented as a directed graph, in which nodes represent components of



Data

| | Exp1 | Exp2 | Exp3 | ... |
|---|---|---|---|---|
| Gene A | 32 | 53 | 11 | ... |
| Gene B | 2 | 34 | 0 | ... |
| Gene C | 34 | 23 | 33 | ... |

**Information Theory Model**
Correlation Coefficients:
A – B = 0.58
A – C = – 0.58
B – C = 1.0

**Boolean Network**
Boolean Functions:
B[t+1] = A[t]
C[t+1] = – A[t] and B[t]

**System of Equations**
Difference Equations:
A[t+1] = 0
B[t+1] = 0.3 · A[t]
C[t+1] = – 0.2 · A[t] + 0.3 · B[t]

**Bayesian Network**
Conditional Probabilities:

| P (B\| A) | A = 0 | A = 1 |
|---|---|---|
| B = 0 | 0.99 | 0.01 |
| B = 1 | 0.01 | 0.99 |

**Figure 8.4** Overview of network models

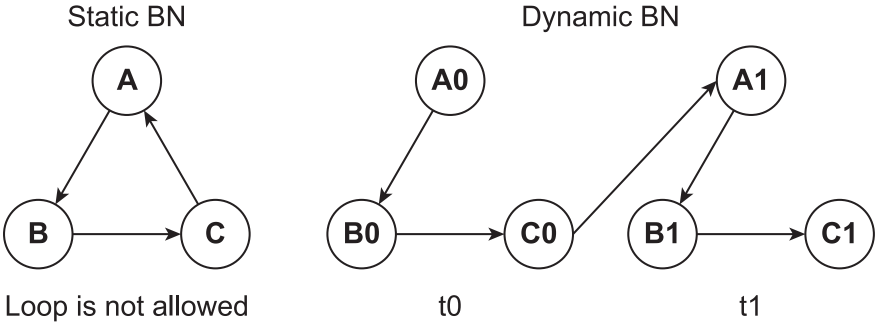


**Figure 8.5** Differences between static and dynamic Bayesian Networks (BNs). A feedback loop from gene A to gene B to gene C and back to gene A is not allowed in static BNs. However, this feedback loop can be represented in a dynamic BN by separating the feedback edges in two time slides.

the system and take one of two discrete values (true or false). Edges between nodes can be represented by Boolean functions made up of simple Boolean operations, e.g. AND, OR, NOT. The Boolean network allows efficient analysis of large regulatory networks. It is relatively easy to interpret, has directed edges, and allows multiple genes to participate in the network, and more importantly, it is dynamic. Boolean networks require the transformation of continuous gene expression signals to binary data. This can be performed, for instance, by clustering and thresholding using support vector regression (Martin et al. 2007). Despite these features, Boolean network is generally criticized because it only allows for two discrete expression levels, clearly an over-simplification of biological processes.

The Differential Equation Model represents changes in gene or protein expression as a function of the expression level of other molecules and environmental factors, and has been widely used to analyze genetic regulatory systems. This model can adequately account for the dynamic behavior of networks by incorporating time-dependent variables, ranging within the set of non-negative real numbers. There are two types of differential equations: linear and non-linear. Linear differential equations can be simply represented as linear algebraic equations. However, the simplification obtained by linearization is not sufficient to identify large-scale networks, and complex dynamic behaviors such as stable oscillatory states cannot be explained using simple linear systems. In contrast, non-linear differential equations can well explain the complicated cellular regulation systems. However, non-linear functions present two major challenges. First, mathematical difficulties are associated with non-linear functions for parameter identification. Second, reliable identification of non-linear interactions normally requires a very large data size. Thus, inference of non-linear systems usually employs predefined functions that reflect prior knowledge to decrease the computational effort needed. But still, the problem of data insufficiency limits the practical relevance of non-linear models. Nevertheless, differential equations have directed edges, allow multiple genes to participate in the regulation and are dynamic, such that they are good candidates for simulating gene regulatory events.

A Bayesian network (BN) represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG), which is a directed graph without feedback loops. The nodes of the graph represent molecular components and its edges represent the causal relationships between molecular components. The relationships are quantitatively encoded in the parameters representing the conditional probabilities (e.g. the probability of a gene being up/down-regulated given the status of other components connecting to the gene). Unconnected nodes represent variables that are conditionally independent of one another. BNs can handle different types of data (e.g. discrete and continuous expression data) and their inference does not require discretization of the data. Nodes in BNs can have multiple parents, such that multiple gene participation is allowed. The approach makes use of the Bayes rule and can be used to reflect the stochastic nature of gene regulation (Werhli & Husmeier 2007). However, the BN is static, the learning process needs relatively large datasets, and the computational cost of the approach is relatively high. Moreover, similar to other network models mentioned above, BNs cannot handle feedback loops, which are an intrinsic feature of many biological systems.

The Dynamic Bayesian Network (DBN) is similar to BN except DBN is able to model dynamic behavior of networks and feedback loops, which occur frequently and are an essential property of many biological systems. DBN adopts the Hidden Markov Model (HMM), a stochastic probability model with hidden variables (Churchill 1989, Rabiner 1989) to model feedback loops by breaking them down into multiple time slices (Figure 8.5.2). DBNs can also handle heterogeneous, incomplete or noisy data (Sun & Hong 2007). Its probability function fits well with the stochastic nature of gene regulation. The drawback of DBNs is that the learning process needs large time-series datasets and the computational complexity is very high.

### 8.5.3 Network Inference

Network inference is achieved through both parameter learning and structure learning. Learning starts with a candidate graph of relationships (a good start will be a graph bearing prior knowledge), followed by parameter learning and structure learning. In parameter learning, the best parameters for each node need to be determined from a given graph and experimental data. And in structure learning, each candidate model is scored according to the graph and the learned parameters. The higher the score, the better the network structure fits the provided data. The final network structure inference result is usually represented by the graph with the highest score, a Bayesian average of multiple graphs, or a distribution of graphs.

### 8.5.4 Challenges in Network reconstruction

There are considerable challenges in computational network reconstruction from biological data. First, the large scale of data from these experiments has inherent variability, as reflected by systematic errors (bias) and stochastic effects (noise) (Hecker et al. 2009). Systematic errors can be nearly eliminated by data normalization. Stochastic effects cannot be completely corrected by data processing, but can be minimized by the application of repeated measurements. Second, many data from biological experiments are incomplete. For example, proteomics data does not contain gene expression information; vice versa. For most biological systems, it is impossible to collect a complete set of data covering every possible measurement. Thus, data integration should be applied to make maximal use of the available data, and the appropriate network models capable of handling incomplete data sets (e.g. DBN) need to be adopted. Third, even for a simple organism, the functional regulation network is complex, as the activity of gene products is regulated by many factors, including transcription factors (TFs) and co-factors that influence transcription, processing of proteins and transcripts, and/or post-translational modification or turn-over of proteins. Moreover, positive and negative feedback add further complexity to the regulation of the network. Finally, the inclusion of large datasets and high degree of network complexity inevitably drive up computational costs. Therefore, depending on the model quality and complexity, the available data and the intended application of identified networks, the suitable model architecture should be carefully chosen in order to efficiently achieve the best results.

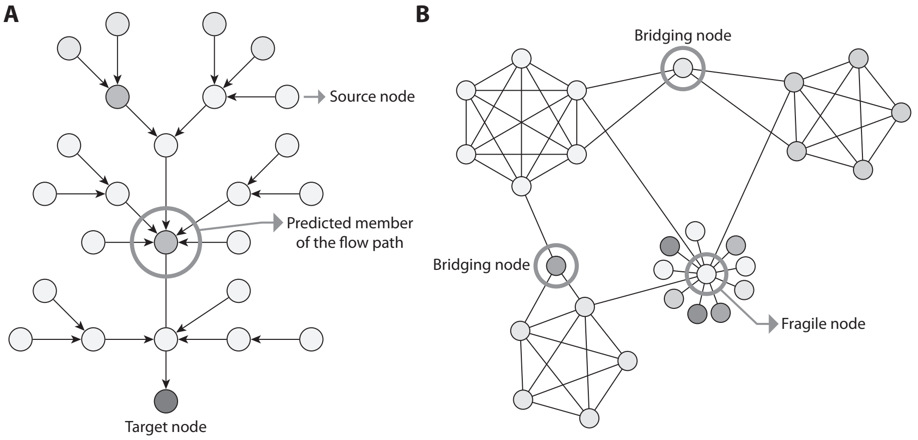


**Figure 8.6** Application of network models. A) Schematic representation of a signal flow path predicted by network modeling (from a directed network). Yellow nodes are source nodes, which are known experimentally to be part of the signal flow. Green nodes are target nodes, where signals converge. The grey nodes are hidden nodes, which are predicted to be part of the signal flow. B) Schematic representation of bridging nodes, which connect two modules. Here, the flow model is applied to an undirected network to distinguish between bridging nodes and fragile nodes (i.e. nodes that have high-betweenness).

## 8.6 Applications of network reconstruction

We expect mathematical modeling of networks to play an important role in generating hypotheses, driving further experimentations and providing novel insights. Some instructive examples come from studies in other organisms. Bonneau et al. were able to reconstruct a significant portion of the regulatory network of the archaeon Halobacterium NRC-1 by integrating genome annotation and gene expression profiles (Bonneau et al. 2006). Several predictions made by the learned network were experimentally tested and verified. Lorenz et al. demonstrated the value of using automatic network inference to identify the regulators of complex phenotypes such as aging (Lorenz et al. 2009). They applied their method to reconstruct interactions in a 10-gene network from the Snf1 signaling pathway, which is required for expression of glucose-repressed genes upon caloric restriction. They also experimentally validated a few predicted interactions, including the demonstration that Snf1 and its transcriptional regulators Hxk2 and Mig1 act as modulators of lifespan. Kaderali et al. developed a Bayesian learning approach that infer pathway topologies from gene knockdown data using Bayesian networks with probabilistic Boolean threshold functions (Kaderali et al. 2009). They demonstrated the power of their results using RNAi data from the Jak/Stat pathway in a human hepatoma cell line. Hong and colleagues reached beyond network reconstruction in single cells and developed a theoretical framework for automatic inference of multicellular regulatory networks (in this case, for *C. elegans* vulval development) by integrating heterogeneous biological data (such as PPIs and gene knockout/knockdown phenotype data) (Sun & Hong 2007, Sun & Hong 2009). The reconstructed model was capable of simulating stochastic *C. elegans* vulval induction under many different genetic conditions, and hence allow researchers to gain systematic view about how animal development is dynamically regulated by interacting cells through complex networks of proteins and genes.

To date, most studies have applied modeling to relatively small networks, centered around one or two pathways. A comparison of two pathway-centered networks can be used to help identify the main routes of pathway cross talk (see review by Hughey et al. (2010)). Simple flow-based models have been used to analyze larger networks. For instance, modeling signal propagation within mammalian hippocampal CA1 neurons revealed global properties of regulation, such as point of signal branching, positions of positive and negative feedback loops within the network (Ma'ayan et al. 2005). An application of network modeling is to go beyond direct observations that can be made from the data and uncover novel components of a cellular response, providing new insights into the biological processes under study. For example, a recent study in yeast integrates genetic perturbation data with protein-protein and protein-DNA interaction networks to predict probable signal flow paths (Huang & Fraenkel 2009, Yeger-Lotem et al. 2009). The model characterized the highest probable flow paths in the PPI network by connecting genetic hits identified from perturbation screens to the corresponding expression changes, revealing novel components within such flow paths. More recently, flow-based network modeling was applied to identification of novel human phospho-ERK modulators (Vinayagam et al. 2011). The flow model used known pERK modulators as source nodes. Hidden nodes downstream of multiple source nodes were predicted to be novel pERK modulators, prediction that was subsequently validated in a cell-based assay. In the context of the *Drosophila* screens described above these approaches now need to be implemented to gain further insights into the structure of the signaling networks.

Predicting novel drug targets is another key application of network models. Network-centered drug-discovery platforms are still in their infancy but some progress has been made. Biological networks are robust in response to removal of most nodes due to redundancy. However, non-redundant nodes appear to be more vulnerable. Network models may facilitate prediction of robust and vulnerable targets based on the network structure. A drug might be effective if it hits a point of fragility in the network; however, targeting an unexpected or extreme point of fragility might lead to more troublesome drug side-effects or toxicity (Figure 6). Thus, the goal for network modeling is to find a set of nodes that are critical in the network structure but at the same time, not so critical that targeting them is likely to lead to global functional impairment (Kitano 2007, Fliri et al. 2009, Schadt et al. 2009). Flow-based models have been proposed to identify bridging nodes (Figure 6), which link two modules. Targeting such nodes only prevents information flow between the modules of interest, not global impairment (Hwang et al. 2008). Recent advancements in developing tools to control complex networks (Liu et al. 2011) will offer a radically new way to develop network based drug targets. It will be exciting to see how increasingly sophisticated and accurate models contribute to our ability to design new avenues of research and gain novel insights into biology.

## ACKNOWELDGEMENTS

S.E.M. is supported by R01 GM067761 from NIGMS. N.P. is an investigator of the Howard Hughes Medical Institute.

## References

Bakal, C. (2011), ‘Drosophila RNAi screening in a postgenomic world’, *Briefings in functional genomics* **10**(4), 197–205.

Bakal, C., Aach, J., Church, G. & Perrimon, N. (2007), ‘Quantitative morphological signatures define local signaling networks regulating cell morphology’, *Science* **316**(5832), 1753–6.

Bakal, C., Linding, R., Llense, F., Heffern, E., Martin-Blanco, E. et al. (2008), ‘Phosphorylation networks regulating JNK activity in diverse genetic backgrounds’, *Science* **322**(5900), 453–6.

Bakal, C. & Perrimon, N. (2010), ‘Realizing the promise of RNAi high throughput screening’, *Developmental cell* **18**(4), 506–7.

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L. et al. (2006), ‘The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo’, *Genome Biology* **7**(5), R36.

Booker, M., Samsonova, A. A., Kwon, Y., Flockhart, I., Mohr, S. E. et al. (2011), ‘False negative rates in Drosophila cell-based RNAi screens: a case study’, *BMC Genomics* **12**, 50.

Boutros, M., Kiger, A. A., Armknecht, S., Kerr, K., Hild, M. et al. (2004), ‘Genome-wide RNAi analysis of growth and viability in Drosophila cells’, *Science* **303**(5659), 832–5.

Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D. et al. (2011), ‘SAINT: probabilistic scoring of affinity purification-mass spectrometry data.’, *Nat Methods* **8**(1), 70–73.

Churchill, G. A. (1989), ‘Stochastic models for heterogeneous DNA sequences’, *Bull Math Biol* **51**(1), 79–94.

Collinet, C., Stoter, M., Bradshaw, C. R., Samusik, N., Rink, J. C. et al. (2010), ‘Systems survey of endocytosis by multiparametric image analysis’, *Nature* **464**(7286), 243–9.

DasGupta, R. & Gonsalves, F. C. (2008), ‘High-throughput RNAi screen in drosophila’, *Methods in molecular biology* **469**, 163–84.

DasGupta, R., Nybakken, K., Booker, M., Mathey-Prevot, B., Gonsalves, F. et al. (2007), 'A case study of the reproducibility of transcriptional reporter cell-based RNAi screens in Drosophila', *Genome biology* **8**(9), R203.

Falschlehner, C., Steinbrink, S., Erdmann, G. & Boutros, M. (2010), 'High-throughput RNAi screening to dissect cellular pathways: a how-to guide', *Biotechnology journal* **5**(4), 368–76.

Fliri, A. F., Loging, W. T. & Volkmann, R. A. (2009), 'Drug effects viewed from a signal transduction network perspective', *Journal of medicinal chemistry* **52**(24), 8038–46.

Flockhart, I., Booker, M., Kiger, A., Boutros, M., Armknecht, S. et al. (2006), 'FlyRNAi: the Drosophila RNAi screening center database', *Nucleic acids research* **34**(Database issue), D489–94.

Friedman, A. A., Tucker, G., Singh, R., Yan, E., Vinayagam, A. et al. (2011), 'Proteomic and functional genomic landscape of receptor tyrosine kinase and Ras/ERK signaling', *Science Signaling***in press**.

Friedman, A. & Perrimon, N. (2006), 'High-throughput approaches to dissecting MAPK signaling pathways', *Methods* **40**(3), 262–71.

Friedman, A. & Perrimon, N. (2007), 'Genetic screening for signal transduction in the era of network biology', *Cell* **128**(2), 225–31.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M. et al. (2006), 'Proteome survey reveals modularity of the yeast cell machinery.', *Nature* **440**(7084), 631–636.

Gilsdorf, M., Horn, T., Arziman, Z., Pelz, O., Kiner, E. et al. (2010), 'GenomeRNAi: a database for cell-based RNAi phenotypes. 2009 update', *Nucleic acids research* **38**(Database issue), D448–52.

Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J.-D. J. et al. (2005), 'Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.', *Nature* **436**(7052), 861–865.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. (2009), 'Gene regulatory network inference: data integration in dynamic models-a review.', *Biosystems* **96**(1), 86–103.

Huang, S. S. & Fraenkel, E. (2009), 'Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks', *Science Signaling* **2**(81), ra40.

Hughey, J. J., Lee, T. K. & Covert, M. W. (2010), 'Computational modeling of mammalian signaling networks', *Wiley interdisciplinary reviews. Systems biology and medicine* **2**(2), 194–209.

Hwang, W. C., Zhang, A. & Ramanathan, M. (2008), 'Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery', *Clinical pharmacology and therapeutics* **84**(5), 563–72.

Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G. et al. (2007), 'Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme.', *Mol Cell* **27**(2), 262–274.

Kaderali, L., Dazert, E., Zeuge, U., Frese, M. & Bartenschlager, R. (2009), 'Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks', *Bioinformatics* **25**(17), 2229–35.

Kaplow, I. M., Singh, R., Friedman, A., Bakal, C., Perrimon, N. et al. (2009), 'RNAiCut: automated detection of significant genes from functional genomic screens', *Nature methods* **6**(7), 476–7.

Kauffman, S. A. (1969), 'Metabolic stability and epigenesis in randomly constructed genetic nets', *Journal of Theoretical Biology* **22**(3), 437–467.

Kitano, H. (2007), 'Biological robustness in complex host-pathogen systems', *Progress in drug research. Fortschritte der Arzneimittelforschung. Progres des recherches pharmaceutiques* **64**, 239, 241–63.

Kockel, L., Kerr, K. S., Melnick, M., Bruckner, K., Hebrok, M. et al. (2010), 'Dynamic switch of negative feedback regulation in Drosophila Akt-TOR signaling', *PLoS genetics* **6**(6), e1000990.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X. et al. (2006), 'Global landscape of protein complexes in the yeast Saccharomyces cerevisiae', *Nature*.

Kummel, A., Gubler, H., Gehin, P., Beibel, M., Gabriel, D. et al. (2010), 'Integration of multiple readouts into the Z-factor for assay quality assessment', *Journal of biomolecular screening* **15**(1), 95–101.

Liu, Y. Y., Slotine, J. J. & Barabasi, A. L. (2011), 'Controllability of complex networks', *Nature* **473**(7346), 167–73.

Ljosa, V. & Carpenter, A. E. (2009), 'Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening', *PLoS computational biology* **5**(12), e1000603.

Lorenz, D. R., Cantor, C. R. & Collins, J. J. (2009), 'A network biology approach to aging in yeast', *Proceedings of the National Academy of Sciences of the United States of America* **106**(4), 1145–50.

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E. et al. (2005), 'Formation of regulatory patterns during signal propagation in a mammalian cellular network', *Science* **309**(5737), 1078–83.

Martin, S., Zhang, Z., Martino, A. & Faulon, J.-L. (2007), 'Boolean dynamics of genetic regulatory networks inferred from microarray time series data.', *Bioinformatics* **23**(7), 866–874.

Mohr, S., Bakal, C. & Perrimon, N. (2010), 'Genomic screening with RNAi: results and challenges', *Annual review of biochemistry* **79**, 37–64.

Niederlein, A., Meyenhofer, F., White, D. & Bickle, M. (2009), 'Image analysis in high-content screening', *Combinatorial chemistry & high throughput screening* **12**(9), 899–907.

Nusslein-Volhard, C. & Wieschaus, E. (1980), 'Mutations affecting segment number and polarity in Drosophila', *Nature* **287**(5785), 795–801.

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F. et al. (2004), 'Multidimensional drug profiling by automated microscopy', *Science* **306**(5699), 1194–8.

Rabiner, L. R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *in* 'Proceedings of the IEEE', pp. 257–286.

Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K., Conaway, R. C. et al. (2008), 'Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics.', *Proc Natl Acad Sci U S A* **105**(5), 1454–1459.

Schadt, E. E., Friend, S. H. & Shaywitz, D. A. (2009), 'A network view of disease and compound screening', *Nature Reviews Drug Discovery* **8**(4), 286–95.

Seinen, E., Burgerhof, J. G., Jansen, R. C. & Sibon, O. C. (2011), 'RNAi-induced off-target effects in Drosophila melanogaster: frequencies and solutions', *Briefings in functional genomics* **10**(4), 206–14.

Shumate, C. & Hoffman, A. F. (2009), ‘Instrumental considerations in high content screening’, *Combinatorial chemistry & high throughput screening* **12**(9), 888–98.

Sims, D., Bursteinas, B., Gao, Q., Zvelebil, M. & Baum, B. (2006), ‘FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets’, *Nucleic acids research* **34**(Database issue), D479–83.

Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. (2009), ‘Defining the human deubiquitinating enzyme interaction landscape.’, *Cell* **138**(2), 389–403.

St Johnston, D. & Nusslein-Volhard, C. (1992), ‘The origin of pattern and polarity in the Drosophila embryo’, *Cell* **68**(2), 201–19.

Sun, X. & Hong, P. (2007), ‘Computational modeling of Caenorhabditis elegans vulval induction’, *Bioinformatics* **23**(13), i499–507.

Sun, X. & Hong, P. (2009), ‘Automatic inference of multicellular regulatory networks using informative priors’, *International journal of computational biology and drug design* **2**(2), 115–33.

van Someren, E. P., Wessels, L. F. A., Backer, E. & Reinders, M. J. T. (2002), ‘Genetic network modeling.’, *Pharmacogenomics* **3**(4), 507–525.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M. et al. (2011), ‘A directed protein interaction network for investigating intracellular signal transduction’, *Science Signaling* **4**(189), rs8.

Walhout, A. J. M., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P. et al. (2002), ‘Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline.’, *Curr Biol* **12**(22), 1952–1958.

Werhli, A. V. & Husmeier, D. (2007), ‘Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge.’, *Stat Appl Genet Mol Biol* **6**, Article15.

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G. et al. (2009), ‘Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity’, *Nature genetics* **41**(3), 316–23.

Zanella, F., Lorens, J. B. & Link, W. (2010), ‘High content screening: seeing is believing’, *Trends in biotechnology* **28**(5), 237–45.

Zhong, W. & Sternberg, P. W. (2007), ‘Automated data integration for developmental biological research.’, *Development* **134**(18), 3227–3238.

Zhu, X., Gerstein, M. & Snyder, M. (2007), ‘Getting connected: analysis and principles of biological networks.’, *Genes Dev* **21**(9), 1010–1024.