A Decade of Systems Biology

Han-Yu Chuang,1,2,* Matan Hofree,3,* and Trey Ideker1−4

1 Division of Medical Genetics, Department of Medicine; 2 Bioinformatics Program; 3 Department of Computer Science and Engineering; and 4 Department of Bioengineering, University of California, San Diego, La Jolla, California 92093; email: tideker@ucsd.edu

Key Words
networks, models, high-throughput technology, genome-wide measurements, developmental biology, cell fate, biomarkers, epistasis, genetic interaction, genome-wide association study (GWAS)

Abstract
Systems biology provides a framework for assembling models of biological pathways from systematic measurements. Since the field was first introduced a decade ago, considerable progress has been made in technologies for global cell measurement and in computational analyses of these data to map and model cell function. It has also greatly expanded into the translational sciences, with approaches pioneered in yeast now being applied to elucidate human development and disease. Here, we review the state of the field with a focus on four emerging applications of systems biology that are likely to be of particular importance during the decade to follow: (a) pathway-based biomarkers, (b) global genetic interaction maps, (c) systems approaches to identify disease genes, and (d) stem cell systems biology. We also cover recent advances in software tools that allow biologists to explore system-wide models and to formulate new hypotheses. The applications and methods covered in this review provide a set of prime exemplars useful to cell and developmental biologists wishing to apply systems approaches to pathways of interest.
INTRODUCTION

Nearly a decade has passed since systems biology was introduced into the language of modern biology (Ideker et al. 2001, Kitano 2002). In that time it has expanded greatly in breadth; it now embraces much of the life sciences and is used to address many research problems across humans and diverse model species (Figure 1). Systems biology has also deepened considerably; many more systematic technologies and methods, both experimental and computational, are in use now than were available a decade ago.

Yet the field in many ways remains in its infancy. The available genome-scale experimental tools are still in an exponential development phase; new technologies turn the field on its head every few years. Even using current assays, bioinformatic methodology lags significantly behind, such that many more data are generated than possibly can be analyzed or interpreted. Moreover, and perhaps most humbling, the field still has not reached consensus on the definition of systems biology. Part of the reason is that systems biology is in vogue, and some have found it easier to change its definition than to change their research habits. However, it is evident that interesting changes are afoot in biology, and given the newness of some these changes, building consensus may take time.

Systems Biology: A Framework for Modeling Biological Systems from Systematic Measurements

Systems approaches, by necessity, involve systematic data. It is impossible to study a biological system as a whole without them. On one hand, the ability to make genome-wide (or proteome-wide or transcriptome-wide) measurements on a system is arguably the single greatest force driving the rise of systems biology. On the other hand, systems biology is not only about genome-scale measurements; it is about a philosophy and a hypothesis-driven approach for experimental design and analysis (Ideker et al. 2001). Therefore, systems biology does not apply to genome-scale studies that are focused solely on discovery. Rather, it is a framework for using genome-scale experiments to perform predictive, hypothesis-driven science (Figure 2). Using genome-scale data to test hypotheses is nontrivial because it requires that the hypotheses themselves be genome-scale. This, in turn, only becomes possible with a genome-scale model of the system. Of course, systematic technologies are not the only means of measuring biological systems. It is critical that systems-level models are consistent with, and validated by, detailed single-molecule measurements and literature.
Figure 1

Meta-analysis of systems biology publications over the past decade. (a) A map of the 20 leading topics in systems biology during the years 2000–2009. The map represents a 2D scaling of the mutual information score between topics, i.e., closely associated topics in the map represent similar themes. The size of the text is roughly proportional to the number of papers. The color gradient indicates a relative increase in citations (from blue to purple to red). Blue indicates topics that were more common prior to 2007; red indicates topics that have been more common since 2007 (see the Supplemental Methods Section for more details on the method and topic word lists). (b) Gray bars show the number of articles indexed in PubMed per year that are labeled with the Medical Subject Heading (MeSH) “Systems Biology.” As a reference, the gold dashed line shows the number of total articles in thousands indexed in PubMed per year.

Enabled by advances in genome-scale technology, the available molecular data are growing exponentially. A property of exponential growth is that the amount of data describing a pathway that will be collected in the next year is on par with the amount of data that has ever been collected about that pathway in the history of science. In light of this fact, clearly the main challenge confronting the field is not to look back (incorporating previous
findings is critical but will be comparatively easy) but to look forward to how one might plan and interpret the mountains of new data that soon will be generated.

Another principle emerging from systems biology research is that it is not enough to map out the physical components and interactions of a system—one must also map how information propagates through this system in response to perturbations. Similarly, it has proven extremely difficult to infer physical or structural interactions in the system from functional data alone (e.g., expression profiles). Thus, systems approaches must necessarily investigate both the physical and functional aspects of the system. For this reason, many approaches seek to integrate multiple data sets, each of which contains a different slice of information about system structure or state.

Finally, as previous authors have done, we distinguish between systems biology and synthetic biology. Systems biology attempts to understand the workings of natural biological systems; synthetic biology uses this understanding to construct new genetic and biochemical systems in vivo or in vitro. Several good reviews of recent progress in synthetic biology are available elsewhere (Andrianantoandro et al. 2006, Benner & Sismour 2005).

A Systems Approach to the Systems Biology Literature

To obtain a systems-level map the current status of the field, we performed a meta-analysis of all systems biology publications recorded in PubMed over the past decade (Figure 1a). The field has grown from a handful of publications published in 2001 to nearly 2,000 published in 2009. Next, we mined the abstracts of systems biology articles published from 2000 to 2009 to extract popular research topics (Figure 1b). We estimated the trends in publication over the decade and compared the topics prominent in systems biology publications prior to 2007 to those in the latter part of the decade (see the Supplemental Methods Section for details; follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org). Certain topics, such as gene expression analyses and evolutionary biology, have maintained their places as mainstays of systems biology (Figure 1a). Others are on the rise, such as cancer research, stem cell, and network biology. A few topics, such as protein structure and comparative genomics, show a decline in publication rates. Nonetheless, the increase in breadth and versatility of research carried out under the banner of systems biology sends a clear message.

In the remainder of this review, we describe progress in systems approaches for mapping biological pathways and for using these maps in biomedical research. Guided by topics in the systems biology meta-map (Figure 1a), we focus on four areas in particular. All of these are strongly emerging topics in systems biology over the past few years: pathway-based biomarkers and diagnosis, systematic measurement and modeling of genetic interactions, systems biology of stem cells, and identification of disease genes. Each of these topics has recently been the focal point of significant research progress brought about because of innovative use of systems-wide measurement methods and computational approaches. In addition, we review the software tools available
for network visualization and interactive exploration of systems biology data, which can be used to formulate hypotheses for further investigation and discovery.

**SYSTEMS APPROACHES TO MOLECULAR DIAGNOSTICS**

A first area in which systems approaches have gained recent traction is in the field of molecular diagnostics. For complex diseases such as cancer, gene and protein expression profiling have become the methods of choice for identifying diagnostic biomarkers able to diagnose the severity of disease and predict future disease outcomes (reviewed by Asyali et al. 2006, Quackenbush 2006, Cheang et al. 2008). Markers are selected by scoring each individual gene or protein on how well its expression pattern can discriminate between different classes of disease or between cases and controls. The disease status of new patients is predicted using classifiers tuned to the expression levels of the markers.

Despite their promise, expression-based diagnostics continue to face serious challenges owing to their questionable accuracy when predicting patient outcomes in some diseases (Ein-Dor et al. 2005, Sotiriou & Piccart 2007). Problems are thought to arise as the result of at least two factors: cellular heterogeneity within tissues and genetic heterogeneity across patients. The impact of cellular heterogeneity depends on the nature of the disease. For some diseases, such as B-cell lymphoma, the diseased cell population is well defined such that it is possible to harvest a relatively pure cell population yielding a distinct expression signature, or to subdivide a mixed B cell population on the basis of expression. In other diseases, such as breast cancer, it has been difficult to cleanly separate tumor from normal cells, such that the resulting expression profile represents an average signal diluted over a mixed cell population.

In contrast, genetic heterogeneity refers to the fact that the same genes may not be dysregulated in each patient. For instance, patient A may have protein A dysregulated, patient B may have protein B dysregulated, patient C may have protein C dysregulated, and so on. Given this disparity across patients who nevertheless may have the same clinical outcomes (e.g., aggressive cancer), classification algorithms have trouble because no single marker is indicative of the status of all (or even most) patients.

To address these problems and improve on expression-based diagnostics, several groups have begun to integrate patient expression profiles with system-wide maps of the pathways in a cell (Anastassiou 2007, Calvano et al. 2005, Doniger et al. 2003, Draghici et al. 2003, Nibbe et al. 2009, Pavlidis et al. 2004, Tian et al. 2005, Ulitsky et al. 2008a, Wei & Li 2007). Depending on the scenario, such pathway maps can involve signaling cascades, transcriptional regulation, or metabolic reactions. They can be as detailed as a series of discrete actions among proteins that lead to a defined end point or functional outcome, or as abstract as a functional annotation on a set of genes. Pathway information provides an overarching layer of organization that can tie seemingly disparate expression responses together into a common pattern. For instance, although any protein A, B, or C may indicate an aggressive form of disease, the knowledge that A, B, and C form a coherent module—e.g., they are subunits of a common protein complex, successive enzymes in a metabolic pathway, or successive steps in a signal transduction cascade—allows us to formulate new biomarker functions that take all of these proteins into account. Some approaches draw this knowledge from known pathways curated from the literature (Subramanian et al. 2005, Vert & Kanehisa 2003); others incorporate pathway knowledge from unbiased networks of physical protein-protein interactions (Chuang et al. 2007, Taylor et al. 2009, Tuck et al. 2006). In either case, the goal is to identify biomarkers not as lists of individual genes or proteins but as functionally related groups of genes or proteins whose aggregate expression accounts for the phenotypic differences between the different populations of patients. Unlike conventional expression
diagnostics based on individual genes, these diagnostic pathway markers provide a strong biological interpretation for the association of an expression profile with a particular type of disease. As a result, the pathway-based approach can be inherently more reliable—which isn’t to say, however, that knowing the pathway relationships assures the success of a diagnostic profile.

In addition to explaining gene expression differences between phenotypes, diagnostic pathways can be used to predict the expression profiles of unknown disease states. Some of these approaches represent pathway activity with a function summarizing the expression values of member genes (Breslin et al. 2005, Guo et al. 2005, Lee et al. 2008); other approaches estimate pathway activation probabilities based on the consistency of changes in gene expression (Efroni et al. 2007, Svensson et al. 2006). Others have engineered normal cells to activate preselected oncogenic pathways to determine gene signatures that can distinguish tumor characteristics (Bild et al. 2006, Glinsky et al. 2005). For example, Bild et al. (2006) overexpressed a panel of oncogenes, one at a time, in primary cultures of human mammary epithelial cells. The goal was to link each oncogene with a distinct set of dysregulated genes. Given these links, they showed that the expression profile of a new tumor sample could be analyzed to identify which oncogenes had been activated.

Chuang et al. (2007) demonstrated an approach that mines pathway biomarkers directly from protein-protein interaction networks. Gene expression profiles of breast cancer patients were superimposed on a human protein-protein interaction network to identify protein subnetworks able to predict cancers likely to metastasize within five years (Figure 3a–c). The activity of a subnetwork was inferred by averaging the normalized expression values of its member genes. The dysregulation of a subnetwork was quantified in terms of the mutual information between subnetwork activity and patient phenotype (metastatic or nonmetastatic). Chuang et al. (2007) also showed that subnetwork markers overlap much more extensively between patient cohorts than individual marker genes and are more informative regarding cancer susceptibility.

Rather than summarizing member gene expression into subnetwork activity, Taylor et al. (2009) proposed to measure changes in interaction coherence between member genes in a subnetwork under different phenotypes (Figure 3d). The interaction coherence in a sample was defined using the difference in expression of the central hub gene in a subnetwork with each of its interacting partners. Although Taylor et al. (2009) and Chuang et al. (2007) differ in the way that they detect pathway dysregulation, both capture a common set of contributions to breast cancer (for example, BRCA1 in Figure 3b versus 3d). Moreover, both studies find that subnetwork markers are more accurate in the classification of breast cancer metastasis than previous predictors based on collections of noninterconnected genes.

In summary, projection of gene expression profiles onto pathway databases or interaction networks is proving to be a powerful approach for understanding disease. On one hand, diagnostic pathways are more reproducible than single genes and can improve the prediction accuracy of disease states. On the other hand, the studies to date are preliminary, and much work is needed before the approach can be translated into advanced diagnostics. One useful direction will be to complement expression and pathway connectivity with other large-scale data sets that include information on genetic perturbations, epigenetic regulation, signal transduction, metabolism, and other factors. Finally, many real and functionally relevant interactions are missing in current protein-protein interaction data sets. Further insights can be expected from reanalysis of the same diseases as the data increase in coverage and quality. Nonetheless, it is clear that constructing functionally coherent, pathway-aggregated biomarkers has great inherent value versus choosing sets of independently-selected genes.
Figure 3
Predictive subnetwork markers for breast cancer metastasis. (a–c) Subnetworks identified by Chuang et al. (2007) involving the key susceptibility regulators (a) TP53, (b) BRCA1, or (c) ERBB2. Nodes and links represent human proteins and protein interactions, respectively. The color of each node scales with the change in expression of the corresponding gene for metastatic versus nonmetastatic cancer. The shape of each node indicates whether its gene is significantly differentially expressed (diamond) or not (circle). The predominant cellular functions are listed next to each module: M, metabolism; CT, cell and tissue remodeling; A, apoptosis; S, signaling of cell growth and survival; CR, cell proliferation and replication. Known breast cancer susceptibility genes are marked by asterisks. (d) BRCA1 and its interactors (e.g., BRCA2 and MRE11, as indicated) are highly ordered (green edges indicate correlated expression between protein pairs) in surviving patients, whereas this organization is lost in patients with aggressive cancer. In contrast, interactions involving SP1 are not significantly altered. PCC denotes the Pearson’s correlation coefficient between the expression patterns of two interacting partners. Panels (a–c) are adapted with permission from Chuang et al. (2007). Panel (d) is adapted with permission from Taylor et al. (2009).
GENETIC INTERACTION MAPS: A TOOLBOX WITH IMPLICATIONS FOR CANCER AND DISEASE

Cell function is governed by a large and complex network of combinatorial interactions among genes, collectively referred to as genetic interactions. Recently, several systems biology studies in yeast, fly, worm, and mammalian cell lines have made important strides in our ability to map this genetic interaction network and its impact on function. Classically, a genetic interaction is defined as the phenomenon whereby combined mutations at several genes produce a phenotype that is unexpected from any of the single mutants (Avery & Wasserman 1992). Genetic interactions are often quantified under the assumption that combining two unrelated (independent) mutations should result in a multiplicative effect on phenotype, such that any deviation is considered an indication of a genetic interaction (Bandyopadhyay et al. 2008, Collins et al. 2006, Costanzo et al. 2010, Dixon et al. 2009, Mani et al. 2008b, St Onge et al. 2007). A phenotypic score that is less than expected is a negative or “aggravating” interaction, whereas a score that is greater than expected is a positive or “alleviating” interaction. An extremely negative genetic interaction that is often studied is the “synthetic lethal” in which the combined gene mutations result in cell death.

When viewed globally over a genome, the network of genetic interactions becomes quite large. To appreciate the magnitude of such a network, consider that among the approximately 30,000 human genes there are on the order of a billion ($30,000^2 = 900,000,000$) potential pair-wise genetic interactions. In a recent near-comprehensive screen of genetic interactions in the yeast *Saccharomyces cerevisiae*, more than 3% of gene pairs showed signs of genetic interaction in rich media conditions (Costanzo et al. 2010). Moreover, genetic interactions need not involve only pairs of genes; rather, they can involve much larger combinations.

Detection of Genetic Interactions in Model Species and Humans

The development of rapid screening techniques for genetic interactions, such as synthetic genetic arrays (SGAs) (Tong et al. 2001, 2004), diploid synthetic lethality analysis by microarray (dSLAM) (Ooi et al. 2003), and epistatic miniarray profiles (E-MAP) (Schuldiner et al. 2005), have allowed the quantification of genetic interaction profiles for the majority of genes in *S. cerevisiae*. Whereas studies of this scope have yet to be implemented in higher organisms, limited genetic interaction screens in human cell lines and model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*, as well as screens in *Schizosaccharomyces pombe*, have already been conducted (Bakal et al. 2008, Bommi-Reddy et al. 2008, Dixon et al. 2008, Lehner et al. 2006, Roguev et al. 2007; for a comprehensive review of epistasis and genetic interaction data sources, see Dixon et al. 2009).

Explicit construction of double gene knockouts in mammals remains a laborious process. Viable alternatives, such as testing combinations of RNA interference (RNAi) knockdowns (Bommi-Reddy et al. 2008, Yang & Stockwell 2008, Zender et al. 2008), are emerging but will naturally take time to mature into genome-scale research tools. In the meantime, a potential role for genetic interaction networks in humans comes from the unlikely direction of statistical genetics, and in particular the technique of genome-wide association study (GWAS). GWAS involves rapidly scanning genetic markers along the genome [such as single nucleotide polymorphisms (SNPs) or copy number variations (CNVs)] to find genetic variations associated with a particular phenotype, such as a heritable trait or disease (Hirschhorn & Daly 2005). However, in many cases GWAS has thus far failed to explain more than a few percent of the genetic contribution to a particular disease, especially for common diseases such as type II diabetes, hypertension, or bipolar disorder (Donnelly 2008, Maher 2008).
Evidence is emerging, however, that some of the missing heritability is attributable to combinatorial genetic interactions within and across pathways (Peng et al. 2009, Torkamani et al. 2008, Wang et al. 2007). The need for inclusion of combinatorial genetic interactions also showcases the importance of developing new approaches to systems-level analysis of genetic interactions (Benfey & Mitchell-Olds 2008).

**Characteristics of Genetic Networks**

Studies in yeast have shown the relative robustness of the cell to systematic deletions, as only a small subset (∼20%) of genes are essential in rich media conditions (Dixon et al. 2009). Hillenmeyer et al. (2008), however, showed that under a variety of stress conditions this list is in fact expanded to include most protein-coding genes (∼97%). As for the effect seen on single deletions, St Onge et al. (2007) reported a ∼24-fold increase in the number of genetic interactions observed after exposure to methyl-methane sulfonate (MMS), a known DNA-damaging agent. Comparisons of normal versus stress conditions suggest that although the genetic network contains some degree of redundancy, it is a highly optimized response mechanism (Costanzo et al. 2010). These experiments illustrate that one should beware of confusing redundancy with robustness.

Topological analysis of the yeast genetic network showed a negative correlation between a gene’s number of genetic interactions and the fitness of its deletion mutant, i.e., hubs in the genetic network tend to have a higher impact on fitness (St Onge et al. 2007). Furthermore, hubs exhibit higher pleiotropy, as estimated by the variety of functional annotations of genetic interactions connected with the hub. A gene’s number of genetic interactions was also found to be correlated with its conservation across yeast species, suggesting that genetic interactions have substantial evolutionary effects (Costanzo et al. 2010). Comparison of genetic interaction networks across different yeasts or between yeast and metazoans suggests that evolutionary conservation is greater at the network level, where the topological characteristics are similar, than at the level of individual interactions, which are not always shared (Dixon et al. 2009, Roguev et al. 2008).

**Integration with Physical Interactions**

Several studies (Bandyopadhyay et al. 2008, Pu et al. 2008, Ulitsky et al. 2008b) have attempted to integrate genetic interaction networks with networks of physical interactions between proteins. As an example, Bandyopadhyay et al. (2008) scored the likelihoods of a protein pair operating either within the same protein complex or between functionally related complexes on the basis of the strength of its genetic and physical interactions. They first learned the appropriate pattern of physical and genetic interactions from known protein complexes curated in databases. Protein pairs with a strong genetic but weak physical interaction typically were found to operate between two functionally related complexes. An agglomerative clustering procedure was then used to merge the protein pairs into increasingly larger complexes and to identify pairs of complexes interconnected by bundles of many strong genetic interactions. Figure 4a shows three example complexes enriched for aggravating genetic interactions (i.e., synthetic lethality).

Hannum et al. (2009) used a similar integrative approach to analyze and reinforce genetic interactions extracted from GWAS (Figures 4b,c). They first identified pairs of SNP markers whose combined state was associated with the expression phenotypes of one or more genes. A biclustering method was then used to discover consecutive intervals of these SNP pairs on two distinct chromosomes and define a genetic interaction network. Similar to Bandyopadhyay et al. (2008), genetic interactions were shown to be strongly enriched within and between known protein interaction complexes. The key difference, however, was that these genetic interactions had been inferred from GWAS rather than generated using directed mutations.
Figure 4

(a) Complexes associated with RAD6-C histone ubiquitination. Protein-protein interactions are enriched among the proteins within each of the three complexes; in contrast, genetic interactions are enriched both within and between complexes. Adapted with permission from Bandyopadhyay et al. (2008). COMPASS, complex of proteins associated with *SET1*; SWR-C, *SWR1* complex; RAD6-C, *RAD6* complex. (b) Two interacting genomic loci (green and blue) that represent significantly dense groups of marker-marker interactions in a genome-wide association study. (c) Interacting complexes spanned by dense bundles of genetic interactions recovered from the same study. Adapted with permission from Hannum et al. (2009).
Emerging Genetic Interaction-Based Approaches to Cancer Therapy

A prominent treatment for cancer is to kill proliferating cancer cells through DNA damage. Because DNA-damaging agents are also toxic to normal tissue, there has been a great deal of interest in developing DNA-damage sensitizers that act specifically on cancer cells via synthetic lethal interactions (Michod & Widmann 2007). In effect, the goal of these studies is to identify and target proteins encoded by genes that are synthetic lethal with cancer-causing mutations. In pioneering work, two groups (Luo et al. 2009, Scholl et al. 2009) have reported promising results from screens focused on finding synthetic lethal relationships with the KRAS oncogene (Figure 5).

Many sensitizers have been or are currently being investigated. Most notably, much attention has been given to a new class of sensitizers known as PARP inhibitors (Farmer et al. 2005). These drugs target an enzyme involved in the base excision repair pathway, which is synthetic lethal with the homologous recombination pathway genes BRCA1 and BRCA2 that are commonly mutated in breast cancer. In addition, farnesyltransferase inhibitors have reached phase III clinical trials, an inhibitor to the cell cycle checkpoint kinase Chk1 is in phase II, and diverse other compounds, such as ataxia telangiectasia mutated (ATM) kinase inhibitors, are under preclinical development. Significant opportunity remains to identify many other potential molecular targets for tumor sensitization, and to date, DNA damage response pathways appear to be a hotbed of such targets. Thus, in addition to the long-term goals of comprehensively mapping the genetic interaction network in different cells under various conditions, the systematic discovery of genetic

Figure 5
A model of mitotic regulation by Ras. (a) BI-2536, a PLK1 inhibitor, attenuates tumor growth in colorectal cancer cells in vivo. Representative images of tumors after treatment are shown. (b) A model in which oncogenic Ras introduces mitotic stress that can be exacerbated to produce lethality by interfering with kinetochore and APC/C (anaphase-promoting complex) function. Genes shaded green are RSL (Regulators of Sex-Limitation) genes, whereas yellow genes cause Ras-specific lethality when overproduced. Red dashed lines illustrate genetic connections between Ras and aspects of mitotic regulation that lead to mitotic stress. Adapted with permission from Luo et al. (2009).
interactions has the potential to profoundly change the treatment of cancer (Mendes-Pereira et al. 2009, Morgan et al. 2010).

SYSTEMS APPROACHES TO IDENTIFY DISEASE GENES

The search for disease-causing genes is a long-standing goal of human genetics. Despite several success stories [e.g., identification of the genetic basis of cystic fibrosis (Rommens et al. 1989), Tay-Sachs (Harding 1983), and Huntington’s disease (Myers 2004)], many diseases with quantifiably substantial genetic components continue to elude detailed genetic explanations (Culverhouse et al. 2002, Moore 2003). For this reason, systems approaches are playing an increasing role in this area through the computational integration of multiple types of genome-wide measurements (Adler et al. 2006, Ergün et al. 2007, Franke et al. 2006, Lage et al. 2007, Mani et al. 2008b, Mullighan et al. 2007, Oti et al. 2006, Tomlins et al. 2005, Yao et al. 2007). Several groups have promoted the idea that similar diseases are caused by mutations in different genes that are part of the same functional module (Goh et al. 2007, Oti & Brunner 2007). The approaches differ in the underlying data sets used, but most of them involve superimposing a set of candidate genes alongside a set of known disease genes on a physical or functional network (Franke et al. 2006, Lage et al. 2007, Oti et al. 2006).

Other methods do not depend on prior knowledge of disease genes but instead infer molecular interaction networks to locate susceptibility genes. For example, Amit et al. (2009) used an RNAi perturbation strategy in mouse dendritic cells to reconstruct the transcriptional network downstream of the Toll-like receptors (TLRS), an important protein family in initiation of pathogen-specific immune responses (Figure 6). Candidate regulators were chosen on the basis of a time course of mRNA expression measured after stimulation with pathogen-derived components. The regulators serve as a gene signature of the immune response in the presence of pathogens. In particular, they identified 144 candidate regulators whose expression changed in response to at least one stimulus. Next, each of the candidate regulators was perturbed (knocked down) by a library of validated lentiviral short hairpin RNAs. Gene expression profiles were gathered under each perturbation and used to infer the regulatory network. The final network included 24 core regulators, affecting the expression patterns of multiple targets, four of which were validated experimentally. They further identified 76 fine-tuners with fewer targets. Together these networks shed light on the regulatory dynamics of the immune response in mammals.

Another approach for de novo identification of disease genes was developed by Wang et al. (2009), who dissected gene expression profiles to infer posttranslational modulators of the MYC transcription factor. Modulators affect a transcription factor at the level of phosphorylation, acetylation, and ubiquitination and are difficult to detect systematically using large-scale methods (Linding et al. 2007). However, key modulators were efficiently identified by computing an information-theoretic measure of correlation between the expression profile of MYC and its direct transcriptional targets, given the expression of a “third party” candidate modulator. Candidate modulators were selected for which the expression level was found to significantly influence the correlation between MYC and its targets. Using a similar information-theoretic measure, Mani et al. (2008a) constructed a network of B cell transcriptional interactions and interrogated it for cancer genes. A similar information-theoretic measure was used to find pairs of interactions in the network that gain or lose correlation when comparing a B cell lymphoma tumor with a reference B cell. They demonstrated that pathways enriched for such high and low correlations may be implicated in pro-oncogenic processes. As a specific example, their method recovered BCL2 and SMADI in follicular B cell lymphoma. Both of these are oncoproteins known to cause cancer but that are not detected through a standard analysis of differential gene expression. Although this approach used expression measurements
Figure 6
A systematic strategy for network reconstruction. (a) Cell state is measured using array-based mRNA expression profiles. (b) From these data, a set of putative regulators is selected. TF, transcription factor; CFs, chromatin modifier factor; RNA bp, RNA-binding protein. (c) The network is perturbed with lentiviral short hairpin RNA (shRNA) against each regulator, followed by measurement of signature genes. (d) These shRNA profiling measurements are used to inform network reconstruction. Adapted with permission from Amit et al. (2009).
and is thus unable to capture the effects of posttranslational regulation, the framework can be easily extended to include measurements of protein level as such high throughput data becomes more commonly available.

Ergün et al. (2007) discovered key mediators in metastatic and nonrecurrent prostate cancers through the use of a regulatory interaction network constructed from a reference set of 1,144 microarray expression profiles spanning seven different cancer types. The known prostate cancer metastasis genes, androgenic receptor (AR) and other genes from the AR pathway, were recovered among the top modulators in metastatic samples but not in non-metastatic ones.

STEM CELL SYSTEMS BIOLOGY AND COMPUTATION OF CELL FATE

Cell fate decisions involve coordinated dynamic expression and regulatory control of hundreds of genes in response to both internal and external stimuli. To dissect the complex interplay among these regulatory pathways, recent studies in stem cell biology have begun to combine classical experimental techniques with emerging high-throughput experimental techniques such as screens for RNAi, genome-wide mRNA expression profiling, large-scale chromatin immunoprecipitation (ChIP), and mass spectrometry-based proteomics (Chen et al. 2008, Kidder et al. 2008, Spooncer et al. 2008). How these vast amounts of data can be used to build a quantitative and predictive model of cell fate control is one of the key challenges in systems biology and stem cell research.

Numerous efforts have been devoted to characterizing the molecular components involved in self-renewal of embryonic stem (ES) cells and differentiation of stem cells along specific lineages. Owing to dramatic advances in genome-wide ChIP technology, the target genes of 20 key ES cell transcription factors, including NANOG (Boyer et al. 2005, Loh et al. 2006, Mathur et al. 2008), OCT4 (Boyer et al. 2005, Loh et al. 2006, Mathur et al. 2008), SOX2 (Boyer et al. 2005), and other factors (Boyer et al. 2005, 2006; Cole et al. 2008; Jiang et al. 2008; Johnson et al. 2008; Kidder et al. 2008; Kim et al. 2008; Liu et al. 2008; Loh et al. 2006; Mathur et al. 2008), have now been identified. An ES cell transcriptional circuit has been assembled through integration of these separate ChIP studies, which cover approximately 50,250 putative protein-DNA interactions that have been identified specifically in ES cells (MacArthur et al. 2009). Moreover, several studies have reported that epigenetic regulation of the key transcription factors by way of chromatin structure (Bernstein et al. 2006, Guenther et al. 2007, Mikkelsen et al. 2007) or DNA methylation (Fouse et al. 2008, Lagarkova et al. 2006, Shen et al. 2006, Yeo et al. 2007) also contributes to the maintenance of pluripotency (reviewed by Bibikova et al. 2008). In addition to epigenetic marks, microRNAs (miRNAs) (Marson et al. 2008) and signaling pathways (Chen et al. 2008) have also been connected to the dynamic balance of ES transcriptional control.

Wang et al. (2006) reported a different take on stem cell systems biology; they assembled a high-quality protein-protein interaction network centered on the \( \text{NANOG} \) transcription factor in mouse ES cells. They used iterative immunoprecipitation experiments to pull down proteins that physically associate with \( \text{NANOG} \), after which mass spectrometry was used to identify the components of the \( \text{NANOG} \) interactome. Interestingly, the \( \text{NANOG} \) interactome is highly enriched in the transcription factors of the core ES cell transcriptional circuit, and many of these factors also regulate the expression of other members of the \( \text{NANOG} \) protein-protein interaction network. This indicates that stem cell fate control is highly dynamic and involves combinatorial interactions between key transcription factors and the genes that encode them. Figure 7 shows the current model of this intrinsically complex but coordinated protein-protein and protein-DNA interaction network.

Müller et al. (2008) reconstructed an extended stem cell regulatory network using a computational approach to integrate publicly
available gene expression profiles and protein interaction networks. They first clustered pluripotent, multipotent, and differentiated human cells on the basis of gene expression and identified a set of genes that are specifically upregulated in undifferentiated pluripotent cells (pluripotency-related genes). Next, using a previously compiled network of human protein-protein and protein-DNA interactions including those in the NANOG interactome (Wang et al. 2006), a collection of subnetworks induced by these pluripotency-related genes were identified using a graph-theoretic algorithm (Ulitsky & Shamir 2007). This collection of subnetworks, which the authors name PluriNet, contains mostly novel interactions; few have been well-characterized in stem cells. Nonetheless, the collection seems to represent common cellular machineries shared by pluripotent cells (including ES cells, embryonal carcinomas, and induced pluripotent cells).

Another recent study has revealed a large map of transcription factor combinations that may point the way to understanding, and perhaps altering, cell fate decisions. Using the mammalian two-hybrid (M2H) system, Ravasi et al. (2010) generated a database of all pairwise protein-protein interactions among the majority (∼1,200) of human transcription factors. From these data, they extracted an interaction network of 15 homeobox transcription factors for which the expression levels were strongly associated with tissue type. The homeobox network was also shown to be capable of stratifying the stem cell expression profiles that had been collected by Muller et al. (2008) into the germ layer from which each was derived (endoderm, mesoderm, ectoderm). It has long been appreciated that combinatorial transcription factor interactions play an important role in cell commitment to different tissue lineages; the work by Ravasi et al. (2010) maps out precisely what some of these combinations are.

Figure 7
Core embryonic regulatory networks for cell fate decisions. (a) High-confidence protein-protein interactions between the transcription factor NANOG and NANOG-associated proteins. An iterative proteomics approach was adapted to identify proteins that physically associate with NANOG and NANOG-associated proteins by using affinity purification in conjunction with mass spectrometry (Wang et al. 2006). (b) Transcription factor binding (protein-DNA) interactions from the data generated by various recent high-throughput chromatin immunoprecipitation (ChIP) experiments. Reproduced with permission from MacArthur et al. (2009).
All of the studies described above support the idea that pluripotency and self-renewal are under tight control by a dynamic and highly complex regulatory network involving protein–protein interactions, transcription factors, signaling pathways, miRNAs, and other epigenetic modifiers. Meanwhile, follow-up experiments are needed to test these inferred regulatory interactions and their effects on stem cell fate. Integration of large-scale RNAi perturbations with genome-wide ChIP experiments and subsequent gene expression profiling (Ding et al. 2009, Hu et al. 2009) has been shown to be useful in confirming a set of transcriptional interactions and their effects on ES cell fate regulation. A next step is to understand how different internal and external stimuli can affect the dynamics of the regulatory network in ES cells. The thorough understanding of such dynamics will enable human-made control of cell fate decisions and, ultimately, tissue engineering and regenerative medicine.

SYSTEMS BIOLOGY SOFTWARE

Considerable time and resources have been expended on developing computational tools for answering systems-level research questions. To effectively analyze systems data, a software tool must meet several requirements. First, it must handle genome-scale data sets. Second, the tool must not be restricted to a single data type but be able to integrate multiple measurements of a system. Third, the software should assist with mapping and modeling of networks and pathways from component data sets. Fourth, it should provide an intuitive interface and visual display of both the data and models.

A number of software packages have been developed to address these requirements. Typically, these packages view the landscape of biological data as belonging to either of two categories: (a) data pertaining to molecular components and their states, and (b) data pertaining to molecular interactions. In what follows, we give a sampling of some of these robust integrative software tools available for systems biology research. Some bioinformatics software is intended for those with an in-depth knowledge of computer science; we focus instead on software tools that are geared toward cell biologists.

Cytoscape is a free bioinformatics environment for integration, visualization, and query of biological networks (Figure 8). Cytoscape’s core software component provides functionality for data import and export, integration of molecular states with molecular interactions, network and integrated data visualization, and data filtering and query tools. Cytoscape’s VizMapper enables attribute-to-visual mappings, which control visual aspects of nodes and edges (e.g., shape, color, size) based on their molecular states (called node attributes). Such mappings allow overlay of multiple data types in a network context.

Cytoscape is developed in Java and disseminated under an open source license (the GNU Lesser General Public License, a permissive software license published by the Free Software Foundation). It has been integrated with many other software tools, including stand-alone applications (e.g., geWorkbench, http://wiki.c2b2.columbia.edu/workbench/index.php), Web sites such as a network image generator (e.g., Harvard Gene Functional Annotation Prediction Browser, http://func.med.harvard.edu/site/yeast/), and major network and pathway databases, including the Biomolecular Interaction Network Database (BIND, http://www.bind.ca/), Reactome (http://www.reactome.org/), the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/), the Michigan Molecular Interactions database (MiMI http://mimi.ncibi.org), and Pathway Commons (http://pathwaycommons.org). Commercial software companies have also used Cytoscape, including Oracle, Agilent GeneSpring and GeneGO (see below).

The Cytoscape core is extended through a straightforward plugin architecture, which allows rapid prototyping and development of advanced computational analyses and features. The active involvement in Cytoscape plugin development by many third-party programmers...
Figure 8
Graphical user interface of Cytoscape. Each window showcases a different analysis or visualization of protein interaction networks and integrated data.

attests to the success of Cytoscape as an open source bioinformatics computing environment. Since 2004 (Cytoscape v2.0–v2.6), more than 74 publicly available plugins have been developed, 46 of which have maintained full compatibility with the latest Cytoscape releases (v2.5 or v2.6) (Cline et al. 2007, Shannon et al. 2003).

NAViGaTOR, another open source network visualization package, is an alternative to Cytoscape. Its use of hardware-based graphics accelerators using Open Graphics Library (openGL) allows fast rendering and visualization of extremely large networks. Interesting options include the ability to visualize graphs using both 2D and 3D views and the ability to collapse nodes into a single “meta node.” NAViGaTOR supports an application programming interface (API) for future plug-ins as well as a variety of data formats. It boasts a lasso selection option and a bookmarking feature to facilitate manual layout and other operations on a network (Brown et al. 2009).

VisANT is a lightweight network visualization tool able to run from a browser-based applet or as a standalone Java program. Of particular interest is its name resolution feature, which attempts to map all nodes in the network to distinct gene names such that two proteins coded by a single gene are always mapped as one entity. This name-mapping feature is one of the most easy to use and streamlined of any software package we review here; it is well designed for the common case with scalability in mind. For large data sets, VisANT has been tested at representing more than 200,000 nodes on a machine with 1 Gb of random-access memory (RAM). Another interesting feature is the
representation of metagraphs, whereby a single node can contain a subgraph. VisANT is also integrated with an online database featuring more than 450,000 interactions in dozens of organisms (Hu et al. 2008).

CellDesigner is a structured diagram editor for drawing gene regulatory and biochemical networks (Figure 9). Users can browse or modify networks as process diagrams (Kitano et al. 2005) and store the networks in systems biology markup language (Hucka et al. 2003), a standard for representing models of biochemical and gene regulatory networks. A unique feature of CellDesigner is that networks are able to link with simulations. Users can view the dynamics of a network under the input parameters through an intuitive graphical interface. CellDesigner is implemented in Java and thus supports various operating systems. The recent releases integrate several simulation/analysis software packages (Funahashi et al. 2008).

Another open-source option is Pathway Assist. The focus of this tool is an automated natural language processing-based information extraction system for protein-protein and gene-gene functional interactions. PathwayAssist also provides a native curated database of protein interactions and cellular pathways. Its text-mining tool can extract biological interactions by reading digital text documents (e.g., biomedical journal articles and abstracts).
efficiently scans sentences, searching for co-occurrences of biological terms and connecting verbs (e.g., the keywords “binds,” “inhibits,” “modulates” or “phosphorylates”) between the co-occurring terms. The bundled database contains at present approximately 500,000 biological interactions among more than 50,000 proteins from several organisms extracted from the current literature (Nikitin et al. 2003). It is available for download upon request.

Finally, two commercial packages are also available—GeneGO (Nikolsky et al. 2005) and Ingenuity (Ingenuity)—both of which offer a comprehensive product aimed at industry and academia.

**SUMMARY AND CONCLUSIONS**

In this review, we have visited four nascent and emerging areas in the field of systems biology. An overarching principle, and one we have tried to highlight throughout, is that systematic measurement techniques coupled with the use of network models lead to the discovery of novel biology. Although one can implement this paradigm in several ways, we have attempted to point out some of the exemplars that have led to big wins in the study of development and disease.

One of the main challenges systems biology will face over the next decade is in breaking the divide between classical and high-throughput methods. Its role is not to replace any of the classical techniques from biochemistry or genetics but to provide a set of organizing principles that integrate these methods (Figure 2). The way forward is undoubtedly through close integration of multiple disciplines to crack the biological system using every means possible.
DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED


