Published in final edited form as:

Curr Opin Biotechnol. 2012 August ; 23(4): 598–603. doi:10.1016/j.copbio.2011.12.005.

Reverse engineering systems models of regulation: discovery, prediction and mechanisms

Justin Ashworth¹, Elisabeth J Wurtmann¹, and Nitin S Baliga^{1,2}

¹ Institute for Systems Biology, Seattle, WA, United States

² Department of Microbiology, University of Washington, Seattle, WA, United States

Abstract

Biological systems can now be understood in comprehensive and quantitative detail using systems biology approaches. Putative genome-scale models can be built rapidly based upon biological inventories and strategic system-wide molecular measurements. Current models combine statistical associations, causative abstractions, and known molecular mechanisms to explain and predict quantitative and complex phenotypes. This top-down ‘reverse engineering’ approach generates useful organism-scale models despite noise and incompleteness in data and knowledge. Here we review and discuss the reverse engineering of biological systems using top-down data-driven approaches, in order to improve discovery, hypothesis generation, and the inference of biological properties.

Introduction

A fundamental goal of systems biology is to reverse engineer the molecular architectures of cellular organisms, by building models to explain, predict, and manipulate their behavior. This includes the ways in which biological systems sense and respond to different internal and external conditions, in order to adapt or produce outputs. This is challenging, as most biological processes are the result of many cellular components involved in many types of interactions. Moreover, the architectures of these interactions and networks are often unknown. New integrative models are needed to connect what is known about the molecular basis of life with complex biological phenomena and the behaviors of whole organisms.

Because our knowledge of biological systems is incomplete, the construction of putative biological models is necessarily based on incomplete information. A new approach to this problem is to adopt the principles of reverse engineering. Reverse engineering is the process by which the essential operation of an unknown system is learned through iterative experimentation and reconstruction. The recapitulation of input and output relationships via a mechanistically inspired model is a primary goal, followed by the prediction of behaviors under new conditions. A suitable operating model can be built with these goals in mind, using all available information. Mechanistic detail and correctness are desirable and beneficial where possible, though not strictly necessary in order to build predictive models.

This kind of approach is well suited to the modeling of under-described biological systems, which includes most (if not all) of those currently under study. ‘Ground up’ models of biology are generally precluded by incompleteness, uncertainty, and the complexity of parameter space — the design and assembly of an entire biological system ‘from the ground

up' is currently far from possible. However, it is possible to query and perturb an existing organism in order to learn about how it functions. This process of reverse engineering involves annotation, data collection, model selection, and strategic perturbations of an existing organism, in order to build a model that is capable of explaining and predicting its current and future behaviors. An example of this data-driven reverse engineering strategy is shown in Figure 1. Below we review current implementations of this approach, as well as new data types and mechanistic considerations that can improve the accuracy and applicability of systems models to scientific and engineering goals.

'Top-down' models capture system-wide properties

Cell-wide molecular measurements, data integration, and pattern-finding algorithms have allowed the construction of comprehensive, predictive models of cellular physiology and regulation for bacterial [1], eukaryotic [2], and archaeal model systems [3]. The success of these models relies on the ability to detect and infer putative interactions and influences based on inherently noisy data and incomplete information. These models have generally relied on a top-down clustering analysis of gene expression under many different environmental and genetic backgrounds. Putative system-wide regulatory states and modules are reverse engineered by grouping genes based on co-expression, shared *cis*-regulatory sequences, and prior information about known associations [4]. Patterns of gene expression, which are the most available data, have generally been used as a proxy for the physiological state of the cell when assessing prediction accuracy. While this represents only one fundamental layer of the biological architecture, it provides a rich target for reverse engineering. Using top-down approaches, the condition-dependent expression of over 80% of genes can be quantitatively predicted as a function of their regulators and the environmental context, even for novel organisms about which little is known *a priori* [3]. The next step is to integrate additional levels of regulation and physiology in order to more directly link regulation to phenotypes. Regulatory and metabolic information are now being combined to model the effects of gene regulation on metabolic and growth phenotypes [5,6].

Systems modeling strives to imply causation and the directionality of interactions, as pattern-finding algorithms alone are insufficient to reflect or reveal the causative mechanisms underlying shared regulatory patterns. Rules governing interactions, transformations, and the flow of information between components are necessary to constrain an otherwise prohibitive solution space. Thus some prior knowledge about the biological 'parts list' and the range of biologically meaningful parameters is essential to constrain model construction. The genome serves as the blueprint to constrain the functions of genes in networks, insofar as it can be completely and accurately annotated [7]. Statistical inference of regulatory relationships has generally involved at least two fundamental mechanistic assumptions: that the expression of a gene or regulon depends on the direct influence of one or more transcriptional regulators, and that these influences operate through direct interaction with shared *cis*-regulatory control sequences in co-regulated genes. This foreknowledge partly addresses the question of 'what is regulating' versus 'what is regulated,' and is fundamental to the inference of causative regulation between a regulators and putative target genes over a range of different cell states [8,9].

These approaches rapidly generate putative regulons, network architectures, and regulatory influences in order to partly explain the systems' inputs and outputs. The architectures of these networks can be shaped by our expectations about network properties, and known examples of regulatory network motifs [10]. It remains difficult, however, to know from such models which associations reflect true physical modules and interactions. The extent to which inferred interaction models are abstractions of true phenomena is often not known, especially for lesser studied organisms. While this does not preclude the statistical prediction

of new behaviors, it is an important limitation for model selection and applicability. Too many hypothetical influences can be drawn between causative agents and their targets, leading to over-fitting and the inference of spurious and/or indirect relationships. New methods exist to automatically reduce model complexity [11,12], partly ameliorating this problem. However, the causative basis of many inferred relationships remains in question until they can be corroborated by additional knowledge or experimentally verified.

Considering the molecular basis of systems biology models

Most observable phenotypes arise by the mechanistic integration of many physically interacting components. To the extent that systems models are compatible with what is known about molecular and biophysical principles, they can also yield a precise basis for experimental hypotheses and the prediction of molecular phenotypes. Each association in a network that models a physical interaction can in theory be experimentally verified or manipulated using molecular techniques. These ‘edges’ in a biological network graph reflect or predict true interactions with mechanistic interpretations. Examples include protein–nucleic acid and protein–protein binding, direct regulatory activity, biochemical reactivity, chemical modifications, metabolic conversions, and dynamics.

Perturbing mechanistically causative components of a system is a powerful means of constraining model selection. In the case of gene regulatory inference networks, transcription factor knockouts correspond to the removal of causative edges, with a clear molecular interpretation — the abrogation of direct interactions that activate or repress transcription [13–15]. Measuring molecular interactions *in vivo* also provides evidence for mechanistic associations. Genome-wide protein–DNA binding, as measured by chromatin immunoprecipitation (ChIP), supports regulatory influences [16] — with caveat that an observation binding (or lack thereof) may not equate directly to function. The addition of nucleosome occupancy [17], chromatin states [18] and the consideration of different activation states of regulators, such as phosphorylation [19**], are further expanding the system-wide mechanistic picture of transcriptional activation and repression.

Model selection can also be made with consideration of post-transcriptional and post-translational regulation, which includes influences from RNA and protein factors acting through a wide range of regulatory mechanisms. As understanding of the role of eukaryotic microRNAs and prokaryotic small RNAs in regulating mRNA abundance has expanded [20], some recent regulatory network models have begun to include these influences [21,22**,23,24]. Moreover, miRNA–mRNA interactions are titrated by competing endogenous mRNAs and non-coding RNAs [25]. A network of such interactions has been inferred for glioblastoma [26**] and regulatory network modeling for other systems should include these interactions. Mechanistic evidence for miRNA-based regulation has come from knockdown and overexpression studies of specific miRNAs and protein cofactors [27,28]. RNA-binding proteins may also exert regulatory influences at multiple levels of gene expression, including termination, 3′ end processing, splicing, export and sub-cellular localization, and translation, that should be included in network modeling. For example, regulatory network modeling of RNA-binding proteins in yeast has demonstrated abundant crosstalk and autoregulation [29**]. Mechanistic support for regulation of mRNAs by RNA-binding proteins has come from measurements of protein–RNA interactions [30]. Global measurement of post-translation modifications including phosphorylation [19], acetylation [31], ubiquitination [32], and protein stability [33] will also allow inclusion of multiple layers of regulation into network modeling. The dynamics of post-translational modification and protein stability are now also being described, such as for global ubiquitination states in the epidermal growth factor response [34]. Thus it is also becoming increasingly possible

to incorporate post-transcriptional and post-translational information in regulatory network models, and these data may increase their accuracy and predictive power.

Biophysical mechanisms inform gene regulatory models

A current challenge in modeling transcriptional regulatory influences is to capture not only the presence or absence of direct regulation of genes by transcription factors, but also the quantitative modulation of expression levels. While many regulatory states operate as bi-stable switches that can be approximated by Boolean networks [35,36], we know from thermodynamics that molecular interactions (and thus functional complexes) form according to hyperbolic or sigmoidal relationships, varying continuously with the abundance of the components [37,38]. The sensitivity of transcriptional activation within this dynamic range is readily observed by measuring gene expression as a function of the concentrations of regulatory factors, and/or their relative binding affinities to promoter binding sites [39[•],40]. Transcription factors also exhibit a continuous range of overlap and variability of their DNA binding specificities [41,42[•]]. Accordingly, the high-throughput characterization of transcription factor specificities is beginning to inform systems models [43]. These continuous and sequence-dependent influence models of activation and repression offer a higher degree of precision and sensitivity. Successful quantitative modeling of gene expression will also benefit from the incorporation of an expanding repertoire of genome-wide measurements that capture post-transcriptional regulation, as cells may use particular layers of gene regulation in particular types of environmental responses [44,45]. Techniques are now available to measure mRNA synthesis and decay rates [46,47] and translation rates [48[•],49].

Another further challenge is to accurately model condition-specific or pattern-specific expression of genes by the combined influence of multiple factors. The conditions under which a gene is transcribed is encoded at least in part by the combinatorial interactions of multiple regulatory factors with *cis*-regulatory elements. A good example of this is the regulatory encoding by enhancers that control early patterning and development in the *Drosophila* embryo, which can be modeled by taking into consideration the binding of multiple different regulatory factors and nucleosomes [37,50]. Similar combinatorial effects can be observed in bacteria [51].

In addition to better reflecting biological mechanisms, systems models are becoming more applicable to new scientific questions and engineering goals. The union of top-down approaches and molecular insights allows the rapid discovery of organization and functions in new biological systems [52–54]. These models, however preliminary and incomplete, provide an essential map of the system as a whole upon which to project information and formulate new hypotheses. For example, the molecular evolution of transcription factors and their binding sites, which is considered a means of rapid evolution and adaptation [55], can begin to be understood and inferred in the context of genome-scale gene regulatory networks. Comparative inferences [56[•]] and homology-based predictions of sequence variation in TF–TFBS interactions [57] may improve the prediction of direct regulation in existing and novel genomes. When coupled to multi-species analyses of putative co-regulatory relationships [58] this may allow the identification of causative variations that can explain or rewire existing gene regulatory networks. As systems models evolve to contain mechanistic information about regulons, promoter architectures, regulatory influences and metabolic pathways, they will provide a basis for forward ‘systems’ re-engineering. Models that integrate both regulatory and metabolic networks can now predict growth phenotypes [5,59], which is a step closer to predicting the complex relationships between regulators and their system-wide effects [60[•]]. Such models will also support the grafting of synthetic modules into existing biological systems [61].

How much detail is optimal and sufficient?

How ‘complete’ does a systems model need to be, and just how precisely does it need to reflect molecular and biophysical details? This question ultimately depends on the availability of data, and the scientific questions, predictions and engineering goals that these models are constructed to address — many of which are molecular in nature. Ultimately, it should be possible to trace and predict the mechanistic flow of information through an entire biological system, from sensing, to signaling, gene regulation, protein production, and ultimately pheno-types. A caveat to the development of mechanistically detailed models is that no practically conceivable model will ever represent all possible physical parameters in a system, nor will enough data ever exist to fully constrain them all. It is also experimentally infeasible to measure, and technically prohibitive to model all possible phenomena in a cell, all possible environmental contexts, and all possible genetic perturbations. This results in uncertainty at multiple levels, to which ‘approximate’ but parsimonious models may be less sensitive [62]. While biological systems are often non-parsimonious [63**], useful biological models will be able to distill the complexity of these systems down to what is influential in any given context, across multiple scales [64]. These reverse-engineered models will then suggest a fewer number of stronger hypotheses, and can be applied to *forward* engineering and the rational, model-based design of new biological functions and outputs.

Acknowledgments

The authors would like to thank Aaron N Brooks and Serdar Turkarslan whose thoughtful reading and comments were helpful in the creation of this article. JA is a Gordon and Betty Moore Foundation Fellow of the Life Science Research Foundation. EJW is NRSA Ruth Kirschstein fellow of the National Institutes of Health. NSB is supported by grants from the NIH (P50GM076547 and 1R01GM077398-01A2), NSF (DBI-0640950 and OCE-0928561) and DOE (DG-FG02-08ER64685, DE SC0004877 and ENIGMA). ENIGMA is a Lawrence Berkeley National Laboratory Scientific Focus Area Program supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Foundational Science through contract DEAC02-05CH11231.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 2007; 5:e8. [PubMed: 17214507]
 2. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet.* 2008; 40:854–861. [PubMed: 18552845]
 3. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, et al. A predictive model for transcriptional control of physiology in a free living cell. *Cell.* 2007; 131:1354–1365. [PubMed: 18160043]
 4. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics.* 2006; 7:280. [PubMed: 16749936]
 5. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA.* 2010; 107:17845–17850. [PubMed: 20876091]

6. Ranganathan S, Suthers PF, Maranas CD. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput Biol.* 2010; 6:e1000744. [PubMed: 20419153]
7. Drew K, Winters P, Butterfoss GL, Berstis V, Uplinger K, Armstrong J, Riffle M, Schweighofer E, Bovermann B, Goodlett DR, et al. The proteome folding project: proteome-scale prediction of structure and function. *Genome Res.* 2011; 21:1981–1994. [PubMed: 21824995]
8. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* 2006; 7:R36. [PubMed: 16686963]
9. Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS One.* 2010; 5:e9803. [PubMed: 20339551]
10. Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet.* 2007; 8:450–461. [PubMed: 17510665]
11. Gustafsson M, Hörnquist M, Lundström J, Björkegren J, Tegnér J. Reverse engineering of gene networks with LASSO and nonlinear basis functions. *Ann NY Acad Sci.* 2009; 1158:265–275. [PubMed: 19348648]
12. Gustafsson M, Hörnquist M. Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge. *PLoS One.* 2010; 5:e9134. [PubMed: 20169069]
13. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet.* 2007; 39:683–687. [PubMed: 17417638]
14. Greenfield A, Madar A, Ostrer H, Bonneau R. DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One.* 2010; 5:e13397. [PubMed: 21049040]
15. Pinna A, Soranzo N, de la Fuente A. From knockouts to networks: establishing direct cause–effect relationships through graph analysis. *PLoS One.* 2010; 5:e12912. [PubMed: 20949005]
16. Gu F, Hsu H-K, Hsu P-Y, Wu J, Ma Y, Parvin J, Huang TH-M, Jin VX. Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Syst Biol.* 2010; 4:170. [PubMed: 21167036]
17. Raveh-Sadka T, Levo M, Segal E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* 2009; 19:1480–1496. [PubMed: 19451592]
18. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature.* 2007; 448:553–560. [PubMed: 17603471]
- 19•. Saleem RA, Rogers RS, Ratushny AV, Dilworth DJ, Shannon PT, Shteynberg D, Wan Y, Moritz RL, Nesvizhskii AI, Rachubinski RA, et al. Integrated phosphoproteomics analysis of a signaling network governing nutrient response and peroxisome induction. *Mol Cell Proteomics.* 2010; 9:2076–2088. [PubMed: 20395639] [These authors used protein mass spectrometry to detect phosphorylation states for yeast transcription factors, and projected these data onto regulatory models to investigate the role of signaling and protein modifications in transcriptional regulatory networks.]
20. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature.* 2010; 466:835–840. [PubMed: 20703300]
21. Su W-L, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. *Mol Syst Biol.* 2011; 7
- 22•. Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ. Functional characterization of bacterial sRNAs using a network biology approach. *Proc Natl Acad Sci.* 2011; 108:15522–15527. [PubMed: 21876160] [A small RNA regulatory network for *E. coli* predicts sRNA targets and regulatory interactions between sRNAs and transcription factors.]
23. Luo Z, Xu X, Gu P, Lonard D, Gunaratne PH, Cooney AJ, Azencott R. miRNA regulatory circuits in ES cells differentiation: a chemical kinetics modeling approach. *PLoS One.* 2011; 6:e23263. [PubMed: 22039400]

24. Plaisier CL, Bare JC, Baliga NS. miRvestigator: web application to identify miRNAs responsible for co-regulated gene expression patterns discovered through transcriptome profiling. *Nucleic Acids Res.* 2011; 39:W125–W131. [PubMed: 21602264]
25. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell.* 2011; 146:353–358. [PubMed: 21802130]
- 26••. Sumazin P, Yang X, Chiu H-S, Chung W-J, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, et al. An extensive microRNA-mediated network of RNA–RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell.* 2011; 147:370–381. [PubMed: 22000015] [This study infers a miRNA-based regulatory network in glioblastoma that includes competing endogenous RNA interactions.]
27. Selbach M, Schwanhaussner B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature.* 2008; 455:58–63. [PubMed: 18668040]
28. Leung AKL, Young AG, Bhutkar A, Zheng GX, Bosson AD, Nielsen CB, Sharp PA. Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol.* 2011; 18:237–244. [PubMed: 21258322]
- 29••. Mittal N, Scherrer T, Gerber AP, Janga SC. Interplay between posttranscriptional and posttranslational interactions of RNA-binding proteins. *J Mol Biol.* 2011; 409:466–479. [PubMed: 21501624] [This regulatory network models post-transcriptional interactions and auto-regulation for RNA-binding proteins in yeast.]
30. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods.* 2011; 8:559–564. [PubMed: 21572407]
31. Choudhary C, Kumar C, Gnäd F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science.* 2009; 325:834–840. [PubMed: 19608861]
32. Danielsen JMR, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, Jensen LJ, Mailand N, Nielsen ML. Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. *Mol Cell Proteomics.* 2011; 10:M110.003590. [PubMed: 21139048]
33. Yen H-CS, Xu Q, Chou DM, Zhao Z, Elledge SJ. Global protein stability profiling in mammalian cells. *Science.* 2008; 322:918–923. [PubMed: 18988847]
34. Argenzio E, Bange T, Oldrini B, Bianchi F, Peesari R, Mari S, Di Fiore PP, Mann M, Polo S. Proteomic snapshot of the EGF-induced ubiquitin network. *Mol Syst Biol.* 2011; 7:462. [PubMed: 21245847]
35. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics (Oxford, England).* 2002; 18:261–274.
36. Davidich MI, Bornholdt S. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One.* 2008; 3:e1672. [PubMed: 18301750]
37. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature.* 2008; 451:535–540. [PubMed: 18172436]
38. He X, Samee MAH, Blatti C, Sinha S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol.* 2010; 6:e1000935. [PubMed: 20862354]
- 39••. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature.* 2009; 457:215–218. [PubMed: 19029883] [These authors used synthetic promoter libraries to train a simple thermodynamic model for promoter-encoded transcriptional activity, showed that it could predict activities for new promoters, and demonstrated that this approach can be used to identify new regulatory connections in the genome.]
40. Amit R, Garcia HG, Phillips R, Fraser SE. Building enhancers from the ground up: a synthetic biology approach. *Cell.* 2011; 146:105–118. [PubMed: 21729783]
41. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. Variation in homeodomain DNA binding revealed by

- high-resolution analysis of sequence preferences. *Cell*. 2008; 133:1266–1276. [PubMed: 18585359]
- 42••. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010; 20:861–873. [PubMed: 20378718] [This demonstrates that precise transcription factor sequence specificities can be rapidly measured at genome scales, revealing patterns of binding that match *in vivo* results. Data of these kinds can be used to guide or constrain the inference of genome-wide direct regulatory network interactions.]
 43. Santos MA, Turinsky AL, Ong S, Tsai J, Berger MF, Badis G, Talukder S, Gehrke AR, Bulyk ML, Hughes TR, et al. Objective sequence-based subfamily classifications of mouse homeodomains reflect their *in vitro* DNA-binding preferences. *Nucleic Acids Res*. 2010 doi: 10.1093/nar/gkq714.
 44. Mehta P, Goyal S, Wingreen NS. A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol Syst Biol*. 2008; 4:221. [PubMed: 18854820]
 45. Shalem O, Dahan O, Levo M, Martinez MR, Furman I, Segal E, Pilpel Y. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol*. 2008; 4
 46. Zeisel A, Kostler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, Rechavi G, Soen Y, Jung S, Yarden Y, et al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol Syst Biol*. 2011; 7
 47. Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, Gnirke A, Nusbaum C, Hacohen N, Friedman N, et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol*. 2011; 29:436–442. [PubMed: 21516085]
 - 48••. Schwanhaussier B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. *Nature*. 2011; 473:337–342. [PubMed: 21593866] [Genome-wide measurement of mRNA and protein abundance and turnover performed in parallel offers insight into the relative contribution of regulation at multiple levels of gene expression.]
 49. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*. 2009; 324:218–223. [PubMed: 19213877]
 50. Kazemian M, Blatti C, Richards A, McCutchan M, Wakabayashi-Ito N, Hammonds AS, Celniker SE, Kumar S, Wolfe SA, Brodsky MH, et al. Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol*. 2010; 8:19.
 51. Kuhlman T, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA*. 2007; 104:6043–6048. [PubMed: 17376875]
 52. Kaur A, Van PT, Busch CR, Robinson CK, Pan M, Pang WL, Reiss DJ, DiRuggiero J, Baliga NS. Coordination of frontline defense mechanisms under severe oxidative stress. *Mol Syst Biol*. 2010; 6:393. [PubMed: 20664639]
 53. Hahne H, Mäder U, Otto A, Bonn F, Steil L, Bremer E, Hecker M, Becher D. A comprehensive proteomics and transcriptomics analysis of *Bacillus subtilis* salt stress adaptation. *J Bacteriol*. 2010; 192:870–882. [PubMed: 19948795]
 54. Konstantinidis KT, Serres MH, Romine MF, Rodrigues JLM, Auchtung J, McCue L-A, Lipton MS, Obratzsova A, Giometti CS, Nealson KH, et al. Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. *Proc Natl Acad Sci USA*. 2009; 106:15909–15914. [PubMed: 19805231]
 55. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. Divergence of transcription factor binding sites across related yeast species. *Science*. 2007; 317:815–819. [PubMed: 17690298]
 - 56•. Dufour YS, Kiley PJ, Donohue TJ. Reconstruction of the core and extended regulons of global transcription factors. *PLoS Genet*. 2010; 6:e1001027. [PubMed: 20661434] [This comparative analysis of the CRP/FNR family of bacterial global regulators identifies shared regulatory features between organisms and sequence-dependent classes of interactions that vary between

them. This kind of analysis provides a strong basis for inferring direct regulatory relationships in new organisms.]

57. Yanover C, Bradley P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* 2011 doi: 10.1093/nar/gkr048.
58. Waltman P, Kacmarczyk T, Bate AR, Kearns DB, Reiss DJ, Eichenberger P, Bonneau R. Multi-species integrative biclustering. *Genome Biol.* 2010; 11:R96. [PubMed: 20920250]
59. Barua D, Kim J, Reed JL. An automated phenotype-driven approach (GeneForce) for refining metabolic and regulatory models. *PLoS Comput Biol.* 2010; 6:e1000970. [PubMed: 21060853]
- 60•. Fendt S-M, Oliveira AP, Christen S, Picotti P, Dechant RC, Sauer U. Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast. *Mol Syst Biol.* 2010; 6:432. [PubMed: 21119627] [These authors examined the condition-specific effects of >100 yeast transcription factor deletion mutants on metabolic flux, and linked this with gene expression, proteomics, and signaling to construct a partial integrated regulatory and metabolic network model for yeast. This is an excellent example of a 'systems' approach that links multiple biological levels into a single model.]
61. Nandagopal N, Elowitz MB. Synthetic biology: integrated gene circuits. *Science.* 2011; 333:1244–1248. [PubMed: 21885772]
62. Davidich M, Bornholdt S. The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model. *J Theor Biol.* 2008; 255:269–277. [PubMed: 18692073]
- 63•. de-Leon SB-T, Davidson EH. Information processing at the foxa node of the sea urchin endomesoderm specification network. *Proc Natl Acad Sci USA.* 2010; 107:10103–10108. [PubMed: 20479235] [This combined experimental and regulatory network analysis of sea urchin development shows convincingly that natural solutions for gene regulation are often non-parsimonious.]
64. Koide T, Pang WL, Baliga NS. The role of predictive modelling in rationally re-engineering biological systems. *Nat Rev Microbiol.* 2009; 7:297–305. [PubMed: 19252506]

Figure 1.

A data-driven reverse engineering strategy for the construction of putative regulatory models, and their application to scientific and engineering goals. Genome-wide clustering of regulatory patterns (top) is combined with large-scale knowledge, measurements and perturbations of direct molecular interactions (bottom) to build putative regulatory models and to infer causative influences (center). This then provides a basis for hypothesis generation and forward predictions, including the rational design of new properties (right).