Genetic network modeling

EP van Someren†, LFA Wessels, E Backer & MJT Reinders

†Author for correspondence Information and Communication Theory Group, Department of Mediametics, Faculty of Information Technology and Systems, Delft University of Technology,, Mekelweg 4, Delft, The Netherlands Tel. +31 15 2786424; Fax. +31 15 2781843; E-mail: E.P.vanSomeren@its.tudelft.nl

The inference of genetic interactions from measured expression data is one of the most challenging tasks of modern functional genomics. When successful, the learned network of regulatory interactions yields a wealth of useful information. An inferred genetic network contains information about the pathway to which a gene belongs and which genes it interacts with. Furthermore, it explains the gene's function in terms of how it influences other genes and indicates which genes are pathway initiators and therefore potential drug targets. Obviously, such wealth comes at a price and that of genetic network modeling is that it is an extremely complex task. Therefore, it is necessary to develop sophisticated computational tools that are able to extract relevant information from a limited set of microarray measurements and integrate this with different information sources, to come up with reliable hypotheses of a genetic regulatory network. Thus far, a multitude of modeling approaches has been proposed for discovering genetic networks. However, it is unclear what the advantages and disadvantages of each of the different approaches are and how their results can be compared. In this review, genetic network models are put in a historical perspective that explains why certain models were introduced. Various modeling assumptions and their consequences are also highlighted. In addition, an overview of the principal differences and similarities between the approaches is given by considering the qualitative properties of the chosen models and their learning strategies.

Keywords: bioinformatics, microarray data analysis, genetic regulatory networks

In pharmacogenomics and related areas, a lot of research is directed towards discovering, understanding and/or controlling the outcome of some particular biological pathway. Numerous examples exist where the manipulation of a key enzyme in such a pathway did not lead to the desired effect [1]. This usually happens because the intended effect was compensated for by the genetic regulation of enzyme levels. Such examples illustrate the importance of accounting for genetic regulation.

We know that the structure of complex genetic and biochemical networks lies hidden in the sequence information of our DNA but it is far from trivial to predict gene expression from the sequence code alone. The current availability of microarray measurements of thousands of gene expression levels during the course of an experiment or after the knockout of a gene provides a wealth of complementary information that may be exploited to unravel the complex interplay between genes. It now becomes possible to start answering some of the truly challenging questions in systems biology. For example, is it possible to model these genetic interactions as a large network of interacting elements and can these interactions be effectively learned from measured expression data?

Since Kauffman [2] introduced the concept of mathematical modeling of complex systems, the reverse engineering of genetic networks has triggered the imagination of many molecular biologists. Somogyi[1] [3] also investigated some of the properties of Boolean networks in relation to biological systems. These researchers showed that Boolean networks possess properties like global complex behavior, self-organization, stability, redundancy and periodicity. Analogies between basins of attraction and different tissue types, as well as cyclic attractors and cell cycles have also been discussed by many other researchers.

Although the behavior and properties of artificial networks match the observations made in real biological systems well, the field of genetic network modeling has yet to reach its full maturity. The automatic discovery of genetic networks from expression data alone is far from trivial because of the combinatorial nature of the problem and the poor information content of

Ashley Publications Ltd www.ashley-pub.com

[1]For reasons of brevity, the authors consistently refer only to the first author of each reference.

the data. First, to model genetic regulation, one needs to take into account the fact that gene expression levels are regulated by the *combined action* of multiple gene products [4]. Second, the number of measurements (arrays) is relatively small compared to the number of measured objects (genes) and the data are corrupted with a substantial amount of measurement noise. Together, these two complicating factors make the construction of genetic networks from empirical observations extremely difficult. In addition, results are further complicated by the presence of inherent noise caused by, for example, variations between different individuals, small numbers of molecules available in a given cell, variations between tissues in a given individual, variations caused by effects that are not measured etc.

The dimensionality problem (many objects and few measurements) plays a fundamental role in genetic network modeling causing the straightforward estimation of model parameters to become extremely unreliable (many equally good solutions). The common approach to avoid this problem is to either reduce the model's complexity or to apply constraints on the parameters. Consequently, the relatively young field of genetic network modeling has been governed by the introduction of a plethora of different models and learning strategies.

This paper provides an overview of genetic network modeling approaches that employ expression data to automatically discover genetic interactions. Developments in this field are placed in a historical context and the qualitative properties and learning strategies of the proposed models are compared and discussed. Recently, another review on genetic network models has appeared [5], but that review focuses more on the mathematical properties of the models.

In this review, the different modeling approaches are first introduced by presenting a historical perspective on the development of genetic network modeling. Then, we briefly consider other approaches that utilize alternative information sources to discover genetic interactions. Focusing on genetic network models with an automatic way of discovering interactions from expression data, we organize them based on the qualitative properties of the underlying model and on their strategy for learning the interactions. Finally, we discuss the current trends in genetic network modeling and indicate which developments are, in our opinion, to be expected in the near future.

The analytic approach: traditional reductionism

For many years, biological research was governed by a reductionist's approach, i.e., a system was investigated by studying the characteristics of its building blocks. Making measurements was laborious and therefore only a handful of elements were typically measured. Consequently, the traditional approach towards genetic network modeling was knowledge driven and based on integrating existing pieces of biological knowledge into a realistic model. Typically, a small complex model was constructed and its parameters tuned manually, until the simulated behavior of the model approached the observations made on the biological system under study. This analytic (or deductive) method of genetic network modeling is still commonly practiced and is currently well-established.

Realistic models of biological phenomena

On the one hand, this approach has led and still leads to very realistic models with manually tuned parameters that can accurately describe some biological phenomena. To name but a few:

- McAdams [6] proposed to use electrical circuits with logical elements to model the circuit diagram of the lysis-lysogeny switch of the bacteriophage lambda in *E. coli*. McAdams also studied the stochastic mechanisms of gene expression in prokaryotes [7].
- Matsuno [8] constructed a hybrid Petri net model (with continuous as well as discrete expression levels) that accurately models the same lysis-lysogeny switch of lambda phage.
- Yuh [9] has constructed a model of the cis-regulatory control of the endo16 gene in the sea urchin, based on sequence information as well as expression data.

A prerequisite for this type of approach is that a lot of *a priori* knowledge is available, which is why it is unsuitable for modeling organisms and pathways about which little information is known. Additionally, one needs to focus on a few genes, thereby possibly ignoring other important factors.

Understanding network behavior

On the other hand, the analytic approach generates a lot of knowledge about the specific properties that an artificial network should possess, such that the simulated expression behavior resembles the real life behavior. Savageau [10], for instance, studied the molecular mode of control

and the capacity of a gene in power-law models in relation to the level of demand. He found that a positive molecular mode of control, where an inducer activates a necessary protein, is best suited for a high level of demand for expression, whereas a negative molecular mode of control, where an inducer removes a repressor-protein, is better suited for a low level. He also considered three types of coupling between regulator and effector genes, i.e., direct coupling, uncoupled and inverse coupling where the regulator gene expression (respectively) increases, stays the same and decreases with an increase in effector gene expression. He found that for a positive mode of control, directly coupling, uncoupled and inversely coupling causes a respectively low, intermediate and high capacity of a gene, i.e., its maximal change in expression. For a negative mode of control this relationship was the exact opposite. Savageau further reasoned that, based on studies of random Boolean networks [11], molecular analysis of bacteria and sequence homology [12], the connectivity of genetic networks is likely to be limited and ranges between 2–12 connections per gene.

Thieffry [13] analytically constructed a qualitative (discrete) model and then studied all of its possible feedback loops. The number of negative interactions inside a feedback loop simply defines the type of circuit, i.e., an even (odd) number defines a positive (negative) circuit. Thieffry found that positive circuits are necessary to obtain multi-stationarity, which is biologically similar to the differentiation of tissues. Negative circuits produce gene profiles that show stable periodicity, a property biologically associated with homeostasis. Studies of known circuits in *E. coli* indicate low connectivity in genetic networks [14], while studies of random Boolean networks indicate small circuit lengths. Together this suggested to him that a genetic network is not a large intertwined network but rather a collection of many small weakly interconnected regulatory modules.

Szallasi [15] defined a 'realistic' Boolean model that consists of nodes representing not only genes but also proteins and their activated states. He remarked that < 2% of the Yeast genes are actually cycling and that no gene changes its expression more than twice during the cell cycle. When simulations were run with his model, however, the number of cycling genes always turned out to be much larger and more frequent changes occurred. This behavior could not be corrected by making changes to some model properties such as the number of ingoing and outcoming connections, the type of canalization, the size of the network or its resulting cycle length. Szallasi further postulated that cancer might be similar to a change from a 'normal' attractor to a 'diseased' attractor. He considered situations in which this change may or may not be reversible. In that sense, the analysis of the basins of attraction for Boolean networks as performed by Wuensche [16] may provide a useful tool.

The notion of a field of attraction is frequently mentioned in the context of Boolean networks but is not specific to the Boolean context. In fact, it originates from the field of continuous dynamic models. Wuensche noted that to be robust and flexible, a biological model must reside on the edge between chaos and order.

Many of the above-mentioned findings are not only interesting for understanding biological processes but may also prove to provide important support for the synthetic approach, i.e., these properties could be incorporated in genetic network models that are learned from expression data.

The synthetic approach: the introduction of large-scale reverse engineering

The introduction of microarray technology made it possible to measure the gene-expression levels of thousands of genes simultaneously. This introduced a new impulse to genetic network modeling, namely the reverse engineering of large-scale genetic networks based on measured expression data. This challenge required a synthetic (or inductive) approach, where a network is constructed automatically, starting from microarray data and a general model of genetic interactions.

This section starts with a historical perspective that describes the introduction of the dynamical models, i.e., models that are learned on time course gene expression data. At the end of this section, the static models that are learned on (steady-state) perturbation gene expression data are described.

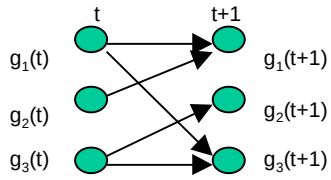
Boolean networks (1998–2000)[2]

In 1998, Liang [17] started off by introducing REVEAL, an algorithm that automatically con-

[2]The indicated years (here and at other headings) refer to the publication year of the papers covered in each section, though in many cases work is still going on.

Figure 1. Example of: a) a Boolean network of three genes with corresponding b) state-transition table and c) Boolean rules.

a) Network:



c) Rules:

$g_1(t+1) = g_1(t)$ AND $g_2(t)$
$g_2(t+1) =$ NOT $g_3(t)$
$g_3(t+1) = g_1(t)$ OR $g_3(t)$

b) State-transition table:

$g_1(t)$	$g_2(t)$	$g_3(t)$		$g_1(t+1)$	$g_2(t+1)$	$g_3(t+1)$
0	0	0		0	1	0
0	0	1		0	0	1
0	1	0		0	1	0
0	1	1		0	0	1
1	0	0		0	1	1
1	0	1		0	0	1
1	1	0		1	1	1
1	1	1		1	0	1

structs a large-scale Boolean network from data. In a general Boolean network model, all gene expression levels are discretized into binary expression levels; a gene is either *on* or *off*. The binary expression levels of all genes in the system at a certain point in time define the state of the network at that time instant. A state transition table defines, for each possible network state, which network state will be next (see **Figure 1b**). From this table, a Boolean rule can be determined for each gene that describes how its expression level at the next time instant depends on some combination of the gene expression levels at the current time instant.

Typical Boolean rules contain logical operators such as AND, OR and NOT (see **Figure 1c**). By placing connections between each of the input genes in the rule and the output gene, the structure of the network can be determined, which expresses the interactions among all genes (see **Figure 1a**). A typical gene expression dataset, after discretization, represents an incomplete state-transition table, since not all possible states will have been measured.

REVEAL constructs the rule for a target gene from this incomplete table by considering the mutual information between the input states of each single gene ($k = 1$) and the output state of the target gene. If the output can be perfectly determined by one of the inputs, the corresponding rules and connections are extracted. If not, all combinations of two genes ($k = 2$) are considered as input and it is examined whether this pair can perfectly predict the target. If not, the procedure repeats for $k = k + 1$ etc. In other words, the structure is learned using a forward exhaustive search procedure that stops as soon as a perfect reconstruction is possible.

A year later, Akutsu [18] proved, using a conceptually simpler approach, that $O(\log_2 N)$ random measurements are sufficient to identify a network of N genes with bounded connectivity K but this algorithm takes $O(N^{K+1}Q)$ time, with Q the number of state transitions. This implies that for a typical gene expression dataset with 1000 genes and connectivity $K = 2$, in the order of 10 independent measurements are sufficient but in that case $O(10^{10})$ time is required! The algorithm learns a Boolean model by performing an exhaustive search not only for each possible combination of inputs but also for each possible configuration of Boolean functions (using only AND or and NOT operators) that are consistent with the given state transitions. Unfortunately, this algorithm was not suited for noisy conditions but a year later Akutsu presented an algorithm that is robust to noise [19,20].

Continuous models (1999–2001)

Although Boolean networks provide a good starting point, they are generally criticized because only two discrete expression levels are allowed. Many examples exist where genes are regulated in a continuous manner rather than

just turned on or off [21-23]. This inspired the introduction of models with a continuous representation of gene expression.

D'Haeseleer [24] learned a linear model on data from the rat central nervous system (CNS), during development and injury after kainate injection [25]. He coupled two partly overlapping datasets, to utilize as much information as possible, resulting in a dataset of 65 genes and 28 time points. Even this simple linear model (with a single parameter per gene) contains more parameters than the number of measurements. This so-called dimensionality problem makes it possible to find many parameter sets that perfectly reconstruct the data. As a result, the parameter estimations become unreliable. To accommodate the fact that the datasets were differently sampled, D'Haeseleer employed a nonlinear interpolation method (resulting in 68 time points). By employing a nonlinear interpolation scheme, he enforces smoothness and tries to avoid the dimensionality problem.

Weaver [26] also employed the linear model but augmented it with a biologically inspired, non-linear dose–response curve. Although nonlinear, this model is essentially a recurrent neural network without a hidden layer. By de-squashing the dose–response curve, the model can be solved by simple linear algebra. To handle the dimensionality problem, Weaver proposed the use of the Moore-Penrose pseudo-inverse. This special matrix inverse produces a solution for underdetermined problems that minimizes the sum of the squared weights but still perfectly fits the data. To introduce limited connectivity, he proposed a greedy backward search that iteratively sets the smallest weight to zero and then recomputes the pseudo-inverse on the now slightly less underdetermined problem. Unfortunately, the de-squashing step is quite sensitive to small changes in the data.

Rather than a discrete-time model, Wahde [27] employed a continuous-time recurrent neural network. A genetic algorithm (GA) was employed to find the parameters of small networks (four genes) learned on the average profiles of clustered data. A genetic algorithm [28] is an optimization technique based on natural selection in which a set of possible solutions, called a population, is evaluated in parallel. New populations of potentially better solutions are generated and evaluated by combining (crossover) and modifying (mutation) the best solutions in the current population. After learning the parameters with a GA, a qualitative description of the parameters is given. Wahde showed results on artificial data as well as on the CNS dataset presented by Wen [25]. Using artificial data he showed that it is better to have multiple shorter time series than one long series. In later work [29,30], he suggested a procedure that forced parameters that were not significant to zero. Repeated elimination of the most unreliable parameters can also be viewed as a form of backward search.

Chen [31] proposed an even more realistic model based on a system of differential equations that models both mRNA and protein levels, including degradation. Chen showed that, provided that both mRNA and protein levels are given, solving this model is similar to the problem of finding minimum weight solutions to linear equations (MWSLE). Unfortunately, this problem is known to be NP complete. However, for a constant connectivity, K, the problem can be solved in $O(QN^{K+1})$ time (using a dataset of N genes and Q time points) by just checking all N^K possible structures. Chen also reasoned that, as many genes showed periodic expression, the Fourier transform for stable systems (FTSS) might be employed as an alternative approach.

A year earlier, Spirov [32] had also suggested the use of a system of differential equations but for a smaller network and with more data points. For learning the parameters, he suggested first using a genetic algorithm to come up with an initial population of globally 'optimal' solutions, which is then used as seeds for a parallel simulated annealing (SA) search. Simulated annealing is a sequential optimization technique that is based on evaluating random changes to the current solution. Better solutions are always accepted, whereas worse solutions are accepted with a probability that decreases during optimization. As a result, SA moves consistently to better solutions but is able to jump out of local optima. When these runs have almost converged, a local gradient descent (GD) approach is employed.

Modeling concerns (1999–2001)

Apart from many papers that introduced a new reverse engineering approach based on yet another model, gradually more papers emerged that addressed the issues associated with genetic network modeling itself. With the reductionist's approach, the combinatorial nature of genetic regulation had largely been ignored [33]. Therefore, it took some time before researchers realized the immense complexity that learning

genetic networks from expression data involved and the early enthusiasm subdued.

Szallasi [33] claimed that there are four factors inherent in biological systems that influence the reverse engineering of genetic networks from expression data. First, the nature of genetic networks is undoubtedly stochastic but microarray measurements are population averaged, which may mask the real individual regulatory interactions. Also, a faster sampling rate is not always possible because the measurement error determines a lower bound on the sampling interval, i.e., the expected difference in expression within one sampling interval should be larger than the measurement noise. Secondly, there are also many regulatory factors that are not modeled, such as (de-)stabilization of mRNA, translocation, phosphorylization etc. Thirdly, he reasons that the information content of the data is not as large as its size would suggest (1–2 orders of magnitude smaller), as only a few genes cycle and even fewer show frequent changes during cell cycle. On the other hand, a property that is favorable for network analysis is that networks are believed to exhibit a high level of compartmentalization.

Spirtes [34] also discussed some of the complicating issues of data acquisition in relation to the construction of genetic networks. Apart from the above-mentioned issues of small sample sizes (dimensionality problem), the substantial measurement error and the masking effect of population averaged measurements, he also points to the fact that the final results can be influenced by hidden (e.g., not modeled) effects and the loss of synchronization of cells.

Erb [35] experimentally examined the influence of measurement noise. He performed Khalil's sensitivity analysis on a complex non-linear model proposed by Mjolsness [67], employing a fully connected network of only three genes. Already with such a small network, the parameters turned out to be very sensitive to noise in the data.

A comparative study done by Wessels [36,37] proposed a set of mathematical properties that genetic network models should possess and by means of which they can be compared. In a small experimental study of continuous models, in which the models were learned on data generated by the other models, he reported disappointing results in terms of how well models can reveal the underlying interactions when faced with noise and limited data. The results favor simple, i.e., linear or pair-wise, models that are less sensitive to unfavorable data conditions[3].

Pairwise models (1997, 1999, 2000)

One way to overcome the dimensionality problem is to restrict the complexity of the model, for example, by only considering pair-wise relationships.

Arkin [38] was the first to suggest the construction of biochemical pathways by means of time-shifted pair-wise correlations. First, the position and magnitude at which the maximal time-shifted cross-correlation occurs is computed in a pair-wise fashion. From this, a distance measure is constructed and single linkage hierarchical clustering is employed, resulting in a singly linked tree that connects associated genes. Augmented with directional and time-lag information this association diagram reveals temporal interactions. Arkin suggested that his approach could also be used to learn genetic networks.

Later, Chen [39] proposed a similar scheme, based on matching peaks in the signals rather than using correlation. After thresholding and clustering, the remaining profiles are represented as a set of peaks. Then peaks in the profiles are compared in a pair-wise fashion to determine causal activation scores. Similarly, inhibition scores are determined. From these scores a putative regulation network is constructed using simulated annealing.

Woolf [40] was the first to describe a fuzzy model for learning genetic interactions. He searched for all possible triplets of an activator and a repressor (two inputs) that influence a target gene (one output). All triplets are scored and ordered on how well they fit the expression data and on whether the inputs showed enough variation.

Unfortunately, these pair-wise (triple-wise) models are fundamentally limited to considering only singly (doubly) connected networks.

Qualitative models (2000, 2002)

A different way to cope with the limitations of the data is to learn qualitative models, thus avoiding the necessity to estimate model parameters precisely.

Akutsu [19,20] described a collection of algorithms that are an intermediate solution, somewhere between Boolean models and continuous differential models. These qualitative models are based on linear differential equations but instead of trying to learn the exact parameters, the

[3]The GA of the Wahde model converged slowly and was therefore stopped early, not allowing the model to converge completely.

researcher derives qualitative abstractions of the parameters. For instance, it is only relevant whether the differences are positive, negative or zero. In this case, a solution can be found by solving a set of inequality relations. Provided that a lot of data are available, these inequalities can be solved using linear programming (LP). Alternatively, the parameters of a non-linear S-system (power-law) can be found using linear algebra by taking the logarithm on both sides of the equations. An S-system is a set of non-linear differential equations of a special form belonging to the power-law formalism (products of exponentially weighted inputs). If the logarithm is taken, the obtained parameter values only portray a relative meaning. But this was exactly the goal: to obtain a qualitative description.

Because of the multitude of detailed biological information acquired over the years, a qualitative model provides an excellent tool to describe the working hypothesis of researchers. Shrager [41] proposed an automatic scheme to revise an initial qualitative model such that it better matches the expression data. This scheme is based on comparing the expected pair-wise correlations of all pairs in the initial scheme with the correlations in the expression data. This measure of data fit is used to construct a fitness function, which is augmented with terms to reduce the number of variables and links in the model. With this fitness function a simple greedy search is performed based on considering single changes in the model. Unfortunately, the employed pair-wise correlation measure does not fully capture the combinatorial nature of the qualitative model.

Modeling revisited (2001–2002)

A better understanding of the consequences of the dimensionality problem resulted in modeling approaches that were better adapted to handle the limitations of the data. For example, strategies started to focus on first reducing the problem (e.g., taking a smaller network, using clustering or structure determination) such that the resulting parameters are estimated more reliably. As a result, the boundaries between the analytic and synthetic approaches gradually became blurred.

Van Someren suggested a number of general approaches to reduce the dimensionality problem by incorporating biologically motivated constraints and showed results from artificial data generated with linear networks. The reduction of the number of genes by clustering gene expression profiles was considered by many [24,27,29,30,32,39,42,46,77]. However, Van Someren [42] studied the relationship between clustering and its effect on the dimensionality problem when learning linear genetic network models. In [43,44], he showed that genetic network models could be made robust to noise by minimizing the first-order derivative of the model's output with respect to its input. For non-linear models, robustness is imposed by learning the model on a set of noisy profiles. To impose limited connectivity of the models, Van Someren [45] compared a number of search algorithms that search for structures with limited connectivity. In this comparison, a forward beam search approach proved to be the best.

Mjolsness [46] also suggested the use of clustered data and learned a system of non-linear differential equations using simulated annealing. Apart from minimizing the prediction error, he included a weight-decay term to minimize the weight values and an exponential term that keeps the parameters bounded in the cost function.

Koza [47] employed genetic programming to determine the structure and rate constants of small metabolic pathways. He showed that it was possible to automatically create a metabolic pathway involved in the phospholipid cycle using 270 time points of E-CELL simulations of a 4-enzyme network where all enzymes were perturbed. Unfortunately, a large amount of data were required.

Maki [48] proposed a two-step approach in which first the structure of a pair-wise Boolean network is learned from the steady-state expressions obtained after perturbation of each gene in the network. The resulting network structure is used to define smaller networks modeled by S-systems. The parameters of these systems are then learned using a GA applied on dynamic data. Unfortunately, this approach still needs a lot of measurements, i.e., at least perturbation experiments of all genes!

Static models (2000–2002)

In parallel with the dynamic models we have described thus far, a different type of model was introduced. Though static models were initially also learned on time course data, they are specifically suited to exploit data from perturbation experiments, such as measured gene expression in steady-state or after knockout of a gene.

Butte [54] presented a simple approach to construct relevance networks. He made pair-wise correlation comparisons of both gene expression

profiles and susceptibility data of anticancer agents. Networks were constructed by keeping only the most significant pairs.

Ideker [55] proposed a simulator–identificator system for perturbation data. His system was especially suited to the experimental cycle in which iterations are made between experimental suggestions and experimental results (he used simulations). The goal was to incrementally determine a static Boolean network using a minimal number of experiments.

In [49], Friedman learned a Bayesian network after gene expressions were discretized into three discrete levels: {-1;0;1}. This paper was based on earlier work [50], where a novel search algorithm was introduced. Confidence measures where estimated on two types of pair-wise features: Markov relations and order relations.Two genes have a Markov relation if there is an edge between them or both are parents of another variable. Order relations are determined based on whether a gene is an ancestor of the other. Results were shown on a selection of 800 genes from the Spellman dataset [51], consisting of 76 microarray measurements of 6177 *S. cerevisiae* open reading frames (ORFs). This dataset represents six different time series' measured under different synchronization conditions. In a similar paper [52], he also considered continuous relations based on linear Gaussian relations.

A year later, Pe'er [53] adapted the Bayesian network framework proposed by Friedman [52] to utilize perturbation data and to specifically construct subnetworks of genes that have high confidence. Subnetworks were constructed using a (forward) greedy hill-climbing method using high-scoring triplets as seeds. This approach was successfully demonstrated on expression profiles of *S. cerevisiae*.

Hartemink [56] also utilized the Bayesian network framework to score different proposed structures. He extended the semantics by augmenting normal edges with annotations that describe positive or negative influences among nodes. The principle was shown on an Affymetrix dataset of 52 arrays of *Saccharomyces cerevisiae* to distinguish part of the galactose pathway. A year later, Hartemink augmented this model with a structure-search algorithm based on simulated annealing [57]. In addition, some interactions were forced to be present based on location data. This new type of data makes it possible to determine where a certain protein binds on the DNA.

Imoto [58] extended Friedman's continuous Bayesian network [52] to cope with non-linear relationships using non-parametric regression. He assumed that a non-linear relationship consists of a sum of simple basis functions. His approach resembles the concept of Fourier decomposition but Imoto used B-splines as basis functions.

An approach specifically suited for perturbation experiments was introduced by de la Fuente [59]. If small perturbations are applied systematically to all genes in a genome, this approach finds a network of regulatory interactions based on co-responses of messenger RNA to a common perturbation.

Yoo [59] suggested an approach that can employ both static observational (non-interventional) expression data as well as static perturbation (controlled expression alteration) data. Bayesian relationships with latent variables were learned in a pair-wise fashion on data presented by Ideker [60].

Considering the number of papers, Bayesian networks seem to be most successful for static expression data. A major drawback of the Bayesian networks is, however, the fact that they cannot handle feedback loops [34]. Feedbacks such as the cell cycle are, however, clearly present in real genetic networks.

Strongly related approaches

In the historical perspective (see section on the synthetic approach), the main focus was on a specific subset of approaches that try to learn networks of genetic interactions. Clearly, there are other methodologies to discover regulatory interactions. These approaches differ in the type of information they discover and the information source they utilize. We distinguish four main approaches:

- spatial-temporal models
- pathway scoring
- promoter analysis
- integrated approaches.

Spatial-temporal models

Spatio-temporal models [61-66] take both spatial and temporal information into account, by modeling the development of a population of interacting cells, each with their own gene expression levels. These models are mostly variants of the model introduced by Mjolsness [67] in 1991. Patterns of gene expression are simulated and compared to patterns measured by *in vitro* fluorescence imaging. In these approaches, the number of considered genes remains limited, partly because of the computational complexity

but mostly because of current restrictions on the measurement techniques.

Obviously, these models cannot employ microarray data because typical microarray measurements represent the average expression level from multiple cell extracts and not of individual cells. From a modeling perspective, however, the spatial-temporal models and the dynamic models have much in common and therefore may benefit strongly from learning from each other's developments.

Pathway scoring

Pathway scoring approaches employ information stored in large databases and (in some cases) match that with measured expression data. Typically, information is used from databases of known pathways, databases with annotations and text databases containing paper abstracts.

Some pathway scoring approaches utilize databases of existing general pathways to enumerate all possible pathway variations and assign scores to each of these potential pathways based on how well they match the measured expression data [56,68]. Scores indicate whether the genes in a pathway were active, coregulated and/or expressed in cascade.

Pavlidis [69] uses no pathway database but scores groups of genes with the same function annotation, based on how well their expression profiles match.

Stephens [70] employs an alternative approach by searching for genetic relations in abstracts of papers stored in Medline. A similar system presented by Wong [71] also includes a tool to visualize the found interactions.

Promoter analysis

Promoter analysis approaches extract regulatory interactions from the sequence information. The upstream regions of genes are searched to determine possible promoter sites. Products of other genes may bind to this promoter sequence and thus can be possible regulators of that gene. To determine these regions, Bussemaker [72] searches for sequences whose occurrence pattern correlates with expression data. Other approaches [73-76] first determine a set of coregulated genes, for example, by means of clustering, and then search the upstream regions for common regulatory motifs.

Integrated approaches

Ideker [60] presented a fully integrated approach on large-scale data in which four main steps were taken:

- define an initial model of a pathway
- perturb components in the pathway and measure the responses in mRNA and protein levels
- check the responses with the model
- refine the model to explain the unpredicted responses.

He was the first to present mRNA expression data (microarrays) as well as protein abundance data, using isotope-coded affinity tag (ICAT) reagents and tandem mass spectrometry (MS/MS) and to integrate this with information from databases of known physical interactions of the galactose pathway.

Clearly the integration of different information sources has been an essential part of the analytical approaches for many years. The main difference, however, is that, thus far, only a limited number of genes were considered. For example, in 1998 Yuh constructed a model of the cis-regulatory control of a *single* gene, i.e., the endo16 gene in sea urchins, using sequence information as well as expression data [9].

Considering the trend toward large-scale approaches that integrate a larger variety of information sources, we might expect that, in the near future, results from pathway scoring and promoter analysis approaches will be integrated within the analytic and synthetic modeling approaches.

Model properties and learning strategies

Thus far, a large number of different modeling methodologies, originating from a plethora of domains, have been proposed for finding regulatory interactions. Among these are Boolean, Bayesian and neural networks, linear and fuzzy models, Petri nets, methods based on ordinary differential equations, Markov models, statistics and cluster analysis. In this section, we will provide a taxonomy based on the principal differences and similarities between the proposed genetic network models and their learning strategies. We restrict ourselves to approaches that present some automatic way of learning the model parameters from gene expression data (i.e., models of section on the synthetic approach). This section first considers the qualitative model properties that are important for choosing an appropriate model. Then the strategies to *learn* the model parameters, once a model has been chosen, are described.

Note that in light of the dimensionality problem, model choice and learning strategy together

Figure 2. A schematic overview of approaches describing a genetic network model with some automatic learning strategy.

Type	Expression represent.	Determ. /stoch.	Relation complex	No. inputs	1997	1998	1999	2000	2001	2002
Static	Continuous	Determ. (pair-wise linear)	Linear	Pair-wise		D'Haeseleer [83]		Butte [54]	Dela-Vuente [59]	Shrager [41]
		Stoch. (combin Bayesian)	Linear	Combin.				Friedman [52]		
			Non-linear	Combin.						Imoto [68]
	Fuzzy	Determ.	Non-linear	Triple-wise				Woolf [40]		
	Discrete (non-linear)	Determ.	Non-linear	Pair-wise		D'Haeseleer [83]				
				Combin. (Boolean)				Ideker [55]		
		Stoch. (combin. Rayesian)	Non-linear				Friedman [50]	Friedman [49,52]	Peer [53] Hartemink [56]	Hartemink [57] Yoo [60]
Dynamic (determ.)	Discrete (combin non-linear Boolean)	Determ.	Non-linear	Combin. (Boolean)		Liang [17]	Akutsu [18]	Akutsu [19,20]		
	Continuous	Determ.	Linear	Pair-wise	Arkin [38]					
				Combin.			D'Haeseleer [24] Chen [31]			
			Non-linear	Pair-wise			Chen [39]			
				Combin.			Weaver [26] Spirov [32]	Wahde [27,29] Akutsu [19,20] Mjolsness [46]	Wahde [30] Koza [47]	
Both	Both	Determ.	Non-linear	Combin.					Maki [48]	

Approaches are ordered vertically by the qualitative properties of the underlying model. When reading this table from left to right, the following properties can be distinguished: 1) dynamic or static model; 2) the expression representation; 3) deterministic or stochastic relationships; 4) relation complexity; and 5) the number of considered input combinations. The year of publication determines the horizontal positioning. A property printed in parenthesis below another property indicates that that property was uniquely determined given the choice of the property printed above. For example, in the first column (determ.) is printed below dynamic. This means that in this table all dynamic models are also deterministic models.

must strike a balance between realism (bias) and parameter reliability (variance). As a consequence, for a fixed dataset, a more realistic/complex model (less bias) will generally result in less reliable parameters (more variance) or require stronger constraints to be applied by the learning strategy (more bias!). The results of several theoretical derivations towards the data requirements of different models under noise-free conditions can be found in [77].

Model properties

As a guide to this section, Figure 2 presents a schematic overview of the chosen taxonomy. In this scheme, approaches are grouped according to a set of qualitative properties that describe the principles that underlie each model.

Static versus dynamic

A principal difference between models is whether static or dynamic relations are modeled. Dynamic models assume that the gene expression levels at

past time instants determine the current (changes in) gene expression levels. Dynamic models generally define a parametric model of interactions and try to estimate the parameters from time course gene expression data. Thus, dynamic models depict dependencies *between* microarray measurements taken at different time instances.

Static models search for causal interactions within microarray measurements that are consistent across multiple microarrays. A nice feature of static models is that they can be applied to time course gene expression data as well as to static data. Interactions are found by searching for mutual dependencies between the gene expression profiles of different genes. Well-known examples are clustering approaches [51,78-81] that group genes that exhibit similar expression levels across a set of experiments. Clustering methods are especially suited for discovering coregulated genes, though clustering should be used with care, i.e., clustering always converges, by construction but it does not always converge to something useful. Clustering methods are very similar to some of the pair-wise static models but serve a slightly different objective. Whereas the clustering methods discover groups of related genes, the pair-wise models select the strongest individual (pair-wise) relationships from the weak [54,82]. Bayesian and Boolean networks are typical examples that allow for more complex static relationships, i.e., they determine whether a gene's expression level can be predicted from a combination of the expression levels of other genes (at the same time instant).

By definition, static and dynamic models represent completely different types of relationships. Unfortunately, in practice, the modeling results tend to be interpreted in the same way.

Gene expression representation

Gene expression measurements are continuous values that represent the relative[4] amount of mRNA copies in a biological sample. Therefore, an obvious choice is to represent the gene expression levels in the model by continuous values. Even more so because many feedback principles require a continuous representation [10].

However, microarray data suffers from a substantial amount of measurement noise. Thus, one might argue that each measured value represents only a qualitative description of the gene expression level (e.g., overexpressed, normal or underexpressed). Therefore, the effective information might be better represented when the data are discretized into a suitable number of discrete levels. Furthermore, a discrete representation is a natural way to adapt the complexity of the model to the quality of the data.

However, designing a discretization method that properly represents the qualitative information in the data is far from trivial. For example, what is the right number of discretization levels? The number of levels chosen should not be too low because this leads to a lot of information being destroyed in the discretization process, whereas too many levels will drastically increase the number of parameters of the model (and consequently decrease the reliability of the estimated parameters).

Boolean models are discrete models with only two gene expression levels (on or off). Consequently, a large discrepancy exists between the measured gene expression signals and their Boolean representation. Though Boolean models are conceptually very useful, some of their dynamic concepts are applicable exclusively within the Boolean context. Thus far, no dynamic Boolean model has been applied to real gene expression data.

Stochastic or deterministic

A deterministic model always predicts the same outcome when the initial conditions are the same. A stochastic model models the probability distribution of possible outcomes. On the cellular level, we know that gene expression is governed by stochastic mechanisms, for example, the binding of RNA polymerase on the promoter site and competition of ribosomes with RNase E for binding on the transcript in prokaryotes [7].

However, microarrays generally measure the total number of mRNA copies across a whole population of cells, causing the data to represent the average of many stochastic effects. Nevertheless, stochastic components are introduced into the data because of measurement noise.

This indicates that stochastic models might be more appropriate. However, although deterministic models do not explicitly model the noise components, this does not mean they cannot handle noisy data. In fact, quite often robustness to noisy data is incorporated implicitly in the learning algorithm. For example, when the least mean square (LMS) algorithm is employed, a Gaussian noise model is implicitly assumed.

[4]The mRNA measurements are relative to each other (or to a reference biological sample). General microarray measurements cannot readily be converted to absolute mRNA copy numbers or mRNA concentrations.

Figure 3. A schematic overview of the learning strategies and constraints employed by the approaches under consideration.

Connectivity	Structure optimization		Subtype	Objective function	Parameter determination (* = analytic)	Paper [Ref.]	Thresholding	Clustering
Fixed	Full connectivity			Min E	Fourier transform	Chen-FTSS [31]		
				Min E	Genetic Algorithm	Wahde [27]		Yes
				Min E	Least mean square*	Akustsu-SSYS1 [15]		
				Min C; E = 0	Least mean square*	Someren [42]	Yes	Yes
				Min E	Metabolic control analysis	DelaVuente [59]		
				Min E	Genetic programming	Koza [47]		
				Min E	GA + SA + GD	Spirov [32]		
				Min (E, SMOOTH)	Least mean square*	D'Haeseleer [24]		
				Min (E, SSW, BND)	Simulated annealing	Mjolsness [46]		Yes
	Limited connectivity		Single	Min E; K = 1	Correlation*	Arkin [38]		
				Min E; K = 1	Correlation*, mutual inf.*	D'Haeseleer [83]		
				Min E; K = 1	Linear programming	Akustsu-QNET1 [19]		
				Min E; K = 1	Linear programming	Akustsu-QNET2 [19]		
				Min E; K = 1	Correlation*	Butte [54]		
				Min E; K = 1	Bayesian scoring metric*	Yoo [60]		
				Min E; K = 1	Peak scoring*	Chen [39]	Yes	Yes
			Double	Min E; K = 2	Fuzzy inference*	Woolf [40]	Yes	
Variable	Search	Greedy	Backward	Min K; E = 0	Psuedoinverse*	Weaver [26]		
				Min (E, K)	Genetic algorithm	Wahde [29,30]		Yes
			Forward & backward	Min (E, K)	BNRC	Imoto [58]		
				Min (E, K)	Correlation*	Shrager [41]		
				Min (E, K)	Bayesian scoring metric*	Friedman [49,50,52]		
		Exaustive	Forward	Min K; E = 0	Boolean lookup table*	Liang [17]		
				Min K; E = 0	Least mean square	Chen-MWSLE [31]		
				Min K; E = 0	Boolean functions*	Akustsu [18]		
				Min K; E < e	Boolean functions*	Akustsu-Bool2 [20]		
				Min E; K ≤ 2	Bayesian scoring metric*	Peer [53]		
			Complete	Min K; E = 0	Boolean lookup table*	Ideker [55]		
				Min (E, K)	Bayesian scoring metric*	Hartemink [56]		
	Optimized together with parameters		Pair-wise search GA	Min (E, K)	N/A	Maki [48]		
			Sim. annealing	Min (E, K)	N/A	Hartemink [57]		

Approaches are ordered vertically by the strategies employed to determine the structure. The connectivity can be either fixed beforehand (full, single or double) or variable and learned from the data. In the latter case, the structure can be determined by means of several search strategies or optimized simultaneously with the parameters. Finally, for each paper, it is indicated which objective function is optimized, how the parameters are determined and whether or not the data was thresholded and/or clustered beforehand. E: Data fit error; K: Connectivity; C: Number of clusters; SMOOTH: Smoothness; SSW: Sum of squared weights; BND: Bound on weights; GA: Genetic algorithm; GD: Gradient descent; SA: Simulated annealing

Interaction complexity

Especially for continuous models, the functional form of the interaction between genes provides a natural way to restrict the complexity of the model. A common choice is to restrict the model to allow only linear relationships. Linear relationships may greatly simplify parameter estimations and in many cases allow analytical (closed-form) solutions. Furthermore, the parameters are relatively easy to understand. However, the linear representation may severely limit the expressive power of the model, for example, many biological systems show non-linear behavior and any aspects of chaotic behavior can only be achieved with non-linear dynamics.

The interaction complexity of Boolean models can also be restricted. This is achieved by allowing only a restricted set of Boolean functions (typically, some combination of AND, OR, NOT and XOR operators).

Input combinations

Finally, we make a distinction between pair-wise models and combinatorial models. Pair-wise models determine relationships between pairs of genes and thus only consider single-gene influences. Combinatorial models allow the combined effect of multiple genes to influence a target gene. A major advantage of the pair-wise approach is that it is less sensitive to the dimensionality problem. This approach needs to consider relatively few relationships (N^2 in total) and requires only a limited amount of measurements (arrays).

However, it is known that gene expression is regulated by the *combined action* of multiple gene products [4] and a major drawback of the pair-wise models is that they do not take this into account.

Trends and observations

In the table in **Figure 2**, some relationships between model properties emerge and approaches that have never been tried become apparent. Read this table from left to right; a property printed in parentheses below another property indicates that the choice of the property on top uniquely determines the property in parentheses. It is not surprising that the choice of a discrete representation generally means that non-linear relationships are considered. Less obvious is the observation that there are no approaches that propose a learning strategy for dynamic stochastic models. Although the possibilities (and the difficulties) have been suggested [34,83], the data requirements would be unfavorable and no implementation has yet been presented for learning dynamic stochastic genetic networks. The deterministic dynamic models employ primarily continuous expression representation and those that are discrete are all Boolean models. The static stochastic models are all Bayesian models. Considering the time line, we see that the reverse engineering of Boolean networks is currently less popular.

A recent trend can be observed to combine different types of expression data. Recent developments include the introduction of models that employ both static as well as dynamic data [47,48]. Hartemink [57], on the other hand, employed location data together with dynamic data.

Learning strategies

The quality of the final result is not only determined by choosing the right model properties. A major issue is to design the inference method (learning strategy) such that the appropriate relations are extracted from the data.

A major difficulty of genetic network modeling is caused by the dimensionality problem, which hampers reliable parameter estimation for the majority of models. Strategies to relieve the dimensionality problem are based on:

- reducing the number of modeled elements (genes)
- increasing the number of samples (microarrays)
- simplifying the model's complexity
- regularization of the inference process.

For the remainder of this section, we assume that the model has been chosen and that its complexity remains fixed.

The main approach to solve the dimensionality problem (apart from simplifying the model) is to utilize additional information (apart from the data) and incorporate constraints on the modeling process. Clearly, it is only sensible to apply constraints if this will produce more (biologically) plausible results.

As a guide to this section, **Figure 3** presents a schematic overview of the considered approaches and the learning strategies and constraints they employ. These items are further explained in the remainder of this section.

Thresholding and clustering

Thresholding and clustering are two strategies that globally reduce the number of modeled elements. Although these strategies are specifically

mentioned as part of the learning strategy in a number of approaches [27,29,30,39,40,42,46], they can be considered as a preprocessing step on the data, rendering them generally applicable to any model. For that reason these strategies are considered optional and are depicted in the last two columns of **Figure 3**.

With thresholding, genes with very small or constant profiles are removed from the data. Thresholding is based on the assumption that such profiles do not exhibit a recognizable regulatory phenomenon and therefore need not be considered. In essence, interactions with thresholded genes are thus assumed to be non-existent.

With clustering, gene expression profiles are grouped together, each group of profiles is replaced by a prototypical profile (e.g., average) and only relations between the prototypical profiles are learned. This procedure assumes that genes with similar expression profiles represent the same regulatory information (are co-regulated) and that their regulatory influences cannot be distinguished from each other and therefore need to be considered as one entity. In essence, the interactions assigned to a prototypical profile represent the integrated effect of all the regulatory influences of the genes in that cluster.

Limited connectivity

The average number of genes that influence the expression of one gene determines the connectivity of the underlying genetic network. On the one hand, it is generally known that gene expression levels are regulated by the *combined action* of multiple gene products [4]. In addition, we know that genetic networks have limited connectivity [10]. Arnone and Davidson conclude from a study of all known cis-regulatory regions that gene expression is influenced on average by between four and eight different gene products [84]. On the other hand, it is likely that the number of connections varies strongly on a per gene basis. Jeong [85] showed that the scaling properties of metabolic networks comply with scale-free networks, i.e., there topology is dominated by a few highly connected nodes which link the rest of the less connected nodes.

Constraining the number of connections is by far the most widely employed constraint and governs the taxonomy indicated in **Figure 3**. The employed learning strategies can be organized in approaches that consider a fixed connectivity or approaches that learn the connectivity from the data. The first category consists of approaches that either do not consider limited connectivity and thus solve for a fully connected network or approaches that consider only one or only two connections per gene (pair-wise and triple-wise models). Approaches of the second category determine the structure either by means of a search procedure or by optimizing the structure and the parameters simultaneously. The search procedure performs iterations between a procedure that suggest a structure and a procedure that determines the parameters.

Figure 3 also depicts which algorithm is used to determine the parameters of the model once the structure of the network is given. The choice of this algorithm strongly depends on the chosen model. To further clarify what each considered approach tries to achieve, we have indicated the objective function corresponding to its goal. This objective function indicates which criteria (e.g., constraints) the approach tries to optimize, as well as the conditions under which it attempts to do so. For example, if the objective function is described by $min\ K;\ E = 0$, this means that this approach tries to minimize the number of connections under the constraint that the data fit error remains zero. Approaches that have a zero data fit error as a constraint are generally not robust to measurement noise. Similarly, $min(E, K)$ means that this approach tries to minimize the connectivity and the error simultaneously. Although this kind of objective is more appropriate to genetic network modeling, the optimal solution to this optimization is not obvious. Clearly, the ordering in the table strongly correlates with the choice of objective function. However, we also observe that a different strategy can be applied to obtain the same goal. For example, the objective function $min(E, K)$ can be obtained by simultaneous optimization of the structure and parameters as well as by means of a search procedure.

Robustness against noise

Apart from constraining the connectivity, there are also other biologically inspired constraints that can be imposed to improve the final results. True genetic networks are, in general, assumed to be robust to noise, i.e., slight distortions of the state generally do not result in completely different behavior. Minimizing the model's first derivative of the output with respect to the input can impose this property. A common approach to achieve this goal is by keeping the parameters within bounds or by minimizing them.

Mjolsness [46] minimizes the data fit error together with the sum of the squared parameters

(SSW) and an exponential function of the parameters (BND). Van Someren [44] has shown that for linear models, minimization of the sum of the squared parameters can also be achieved by learning the model on a dataset augmented with noise-corrupted duplicates. Such an approach would provide a simple alternative to parameter optimization for non-linear models

Other constraints

Time course gene expression profiles generally exhibit smooth changes over time. D'Haeseleer [24] employed this property to augment the number of measurements by means of non-linear interpolation (see SMOOTH in **Figure 3**). However, the information gained by additional sampling along a given trajectory is rather low and the process of interpolation to increase the dataset size should be used with great care.

There exist a number of other general properties that have not been implemented as a constraint in any genetic network modeling approach so far. Genetic networks are assumed to be redundant, i.e., cells maintain homeostasis and are found to be robust against many mutations [1]. As scale-free networks [85] are known to be robust and error-tolerant, perhaps genetic networks with a similar topology should be favored. Genetic networks must also be stable systems, i.e., the amount of mRNA of any gene must remain finite at all times. Furthermore, genetic networks are assumed to be highly compartmentalized, i.e., the network consists of compartments with many connections within a compartment but few connections between compartments [33,14].

Outlook and expert opinion

Successful genetic network modeling may prove to become one of the most promising tools of twenty-first century functional genomics. However, the modeling approaches discussed in this review have yet to proof their ability to extract substantial new knowledge about genetic interactions. Before genetic network modeling will grow into a tool for biologist that is used on a regular basis, more efforts are necessary to improve the understanding of model differences, improve model performance and to set up a proper validation test-bed.

Over the years, many different models have been proposed for the discovery of genetic interactions among genes. Although one might have expect that the most suited model would eventually prevail, we are currently aware that, though the field is still young, in 4 years of research we have still have no indication which model is the most suited to genetic network modeling. Of course, the obvious remark here should be that each model reveals a different kind of information and that each is valuable in its own respect. But still, we do not (and perhaps never will) really understand how to handle the differences in results obtained by each different model.

For example, it is not easy to understand how to interpret the differences in results between static and dynamic models. Basically, the main difference in learning these models is whether or not the output data are shifted in time with respect to the input. It is clear that static relations tell us something about common regulations among genes, whereas time relations convey information about which expression effects follow each other in time. We believe that a thorough investigation of how these different results should be related to known biological interactions may prove to be very beneficial.

At least for dynamic relationships, it is obvious that if the complexity of the model is increased sufficiently, the model will describe the underlying complex biological processes more accurately. Clearly, when only a fixed amount of measurements is available, more parameters will generally mean less reliable estimates. Thus, it is necessary to make a trade-off between accuracy (bias) and uncertainty (variance). However, as we have indicated, sensible incorporation of constraints eventually provides a way to allow complex models to be learned even with a limited amount of data.

The modeling trend that is revealed by this review is the use of a larger variety of information for learning genetic network models, be it in terms of other types of measurements, information stored in databases or desired properties of networks. We advocate this trend and encourage dynamical modelers in particular to integrate the network properties discovered by analytical approaches. Similarly, analytical approaches should benefit from automatic procedures to learn parameters. From a broader perspective we would advocate the use and integration of all types of additional information and measurements that are available in databases.

We also found that few researchers [26,27,29,30,42-45] validate the proposed learning strategy on artificial data. Of course, it is essential to test the final approach on real data but in the absence of a ground truth, the performance of an algorithm cannot be evaluated. If the

Highlights

- The concept of mathematical modeling of complex systems, in the context of the reverse engineering of genetic networks from microarray data has great potential to unravel the complex interplay between genes through a systems approach.
- This review shows how a plethora of different genetic network models have been introduced, each in answer to the shortcomings of the models present up to that time.
- The first prerequisite to solve the not-trivial task of genetic network modeling is to have a rich dataset coming from a variety of sources, such as different measurement conditions (static as well as dynamic microarray data), location data and databases.
- The second prerequisite to solve this problem is that the modeling should exploit the general network characteristics, such as limited connectivity, network compartmentalization and feedback principles. These properties can be drawn from already existing biological knowledge and simulation studies.
- Partly since this research field is relatively young, none of the model approaches thus far have revealed substantial new knowledge. To turn genetic network modeling into a reliable tool, 1) model differences should be better understood; 2) model performance and reliability should be increased; and most importantly 3) a proper (biological) validation test-bed needs to be set up.
- The current trend of integrating different sources of information supplies all the elements for a fully integrated approach towards genetic network modeling. Therefore genetic network modeling is bound to bring a new impulse to the field of pharmacogenomics in the coming years.

results are poor, we still do not know whether the learning strategy has failed to find reliable interactions, whether the model did not express the real phenomenon, whether the data were wrongly preprocessed or whether the limited measurements just did not contain sufficient information. Along the same line of thought, one should relate any results found by an approach to the chance of finding anything at random, i.e., proper confidence measures should be developed.

To boost confidence in genetic network modeling and to improve cross-hybridization between model developments, it would be very fruitful to, on the one hand, develop a good set of benchmarking tools to thoroughly investigate the reliability of proposed approaches on artificial data. On the other hand, a well-studied pathway of a familiar organism with a diverse set of measurements should be used as a general test bed for biological validation.

We have witnessed the highly valued introduction of large-scale reverse engineering, followed by increased skepticism when quick results did not follow. Much progress has been made since the fundamental problem was realized and the full complexity of the task was understood. We believe that genetic network modeling is on the verge of proving its potential and in the coming years will provide a new impulse to the field of pharmacogenomics.

Acknowledgements

This work was funded by the DIOC-IMDS programme of the Delft University of Technology. We wish to thank the three anonymous reviewers for their useful suggestions.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. BAILEY JE: Lessons from metabolic engineering for functional genomics and drug discovery. *Nature Biotechnol.* 17(7), 616-618 (1999).
2. KAUFFMAN SA: The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press. Oxford University Press (1993).

• **Describes important properties of Boolean networks.**

3. SOMOGYI R, SNIEGOSKI CA: Modeling the complexity of genetic networks: understanding multigene and pleiotropic regulation. *Complexity* 1, (45) (1996).
4. HOLSTEGE FC, JENNINGS EG, WYRICK JJ *et al.*: Dissecting the regulatory circuitry of a eukaryotic genome. *Cell.* 95(5), 717-728 (1998).
5. DE JONG H: Modeling and simulation of genetic regulatory systems: a literature review. *J. Comp. Biol.* 9(1), 67-103 (2002).

•• **A mathematical overview of genetic network models.**

6. MCADAMS HH, SHAPIRO L: Circuit simulation of genetic networks. *Science*. 269, 650-656 (1995).
7. MCADAMS HH, ARKIN A: Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94, 814-819 (1997).
8. MATSUNO H, DOI A, NAGASAKI M, MIYANO S: Hybrid petri net representation of gene regulatory network. *Pacific Symposium on Biocomputing 2000*. World Scientific Publishing Co. 5, 338-349 (2000).
9. YUH CH, BOLOURI H, DAVIDSON EH: Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896 (1998).

• **Detailed example of the regulation of a single gene.**

10. SAVAGEAU MA: Rules for the evolution of gene circuitry. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 54-65 (1998).
11. KAUFFMAN SA: Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theoret. Biol.* 22, 437-467 (1969).
12. OTSUKA J, WATANABE H, MORI KT: *J. Theoret. Biol.* 178, 183-204 (1996).
13. THIEFFRY D, THOMAS R: Qualitive analysis of gene networks. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 77-88 (1998).
14. THIEFFRY D, HUERTA AM, PEREZ-RUEDA E, COLLADO-VIDES J: From specific gene regulation to global regulatory

networks: A characterization of *Escherichia coli* transcriptional network. (submitted).
15. SZALLASI Z, LIANG S: Modeling the normal and neoplastic cell cycle with realistic boolean genetic networks: their application for understanding carcinogenesis and assessing therapeutic strategies. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 66-76 (1998).
16. WUENSCHE A: Genomic regulation modeled as a network with basins of attraction. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 89-102 (1998).
17. LIANG S, FUHRMAN S, SOMOGYI R: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 18-29 (1998).
18. AKUTSU T, MIYANO S, KUHARA S: Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 17-28 (1999).
19. AKUTSU T, MIYANO S, KUHARA S: Algorithms for inferring qualitative models of biological networks. *Pacific Symposium on Biocomputing 2000.* World Scientific Publishing Co. 5, 290-301 (2000).
20. AKUTSU T, MIYANO S, KUHARA S: Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* 16, 727-734 (2000).
21. RUVKUN G, GIUSTO J: The *Caenorhabditis elegans* heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. *Nature* 338, 313-319 (1989).
22. KATZ WS, HILL RJ, CLANDINI TR, STERNBERG PW: Different levels of the *C. elegans* growth factor LIN-3 promote distinct vulval precursor fates. *Cell*, 82, 297-307 (1995).
23. KALTHOFF K: Analysis of biological development. McGraw-Hill (1996).
24. D'HAESELEER P, WEN X, FUHRMAN S, SOMOGYI R: Linear modeling of mrna expression levels during cns development and injury. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 41-52 (1999).
25. WEN X, FUHRMAN S, MICHAELS GS *et al.*: Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* 95 (1), 334-339 (1998).
26. WEAVER DC, WORKMAN CT, STORMO GD: Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 112-123 (1999).
27. WAHDE M, HERTZ J: Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems* 55, 129-136 (2000).
28. MITCHELL M: An introduction to genetic algorithms. MIT Press, Cambridge (1996).
29. WAHDE M, HERTZ J: Modeling genetic regulatory dynamics in neural development. *J. Computational Biology* 8, 429-442 (2000).
30. WAHDE M, HERTZ J, ANDERSSON ML: Reverse engineering of sparsely connected genetic regulatory networks. *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology* (2001).
31. CHEN T, HE HL, CHURCH GM: Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co4, 29-40 (1999).
32. SPIROV AV *et al.*: The inverse problem for ode models of genetic networks: the determination of parameters of the system from experimentally observed data. *http://academic.mssm.edu/molbio/tmp/circuits/box1circ.html* (1998).
33. SZALLASI Z: Genetic network analysis in light of massively parallel biological data acquisitions. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 5-16 (1999).
34. SPIRTES P, GLYMOUR C, SCHEINES R: Constructing bayesian network models of gene expression networks from microarray data. *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology* (2000).
35. ERB RS, MICHAELS GS: Sensitivity of biological models to errors in parameter estimates. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 53-64 (1999).
36. WESSELS LFA, VAN SOMEREN EP, REINDERS MJT: A Comparison of genetic network models. *Pacific Symposium on Biocomputing 2001.* World Scientific Publishing Co. 6, 508-519 (2001).
37. VAN SOMEREN EP, WESSELS LFA, REINDERS MJT: Genetic network models: a comparative study. *Proceedings of SPIE, Micro-arrays: Optical Technologies and Informatics*, San Jose, California: 236-247 (2001).
38. ARKIN A, SHEN P, ROSS J: A test case of correlation metric construction of a reaction pathway from measurements. *Science* 277, 1275-1279 (1997).
39. CHEN T, FILKOV V, SKIENA S: Identifying gene regulatory networks from experimental data. *Proceedings of the third annual international conference on Computational molecular biology (RECOMB99).* Association for Computing Machinery 94-103 (1999).
40. WOOLF PJ, WANG Y: A fuzzy logic approach to analyzing gene expression data. *Physiological Genomics* 3, 9-15 (2000).
41. SHRAGER J, LANGLEY P, POHORILLE A: Guiding revision of regulatory models with expression data. *Pacific Symposium on Biocomputing 2002.* World Scientific Publishing Co. 7, 486-497 (2002).
42. VAN SOMEREN EP, WESSELS LFA, REINDERS MJT: Linear modeling of genetic networks from experimental data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology.* AAAI, La Jolla, California (2000), 355-366.
43. VAN SOMEREN EP, WESSELS LFA, REINDERS MJT, BACKER E: Robust genetic network modeling by adding noisy data. *Proceedings of the 2001 IEEE – EURASIP Workshop on Nonlinear Signal and Image Processing.* Baltimore, Maryland (2001).
44. VAN SOMEREN EP, WESSELS LFA, REINDERS MJT, BACKER E: Regularization and noise injection for improving genetic network models. Chapter 12 in *computational and Statistical Approaches to Genomics*. Kluwer (2002) :211-226.
45. VAN SOMEREN EP, WESSELS LFA, REINDERS MJT, BACKER E: Searching for limited connectivity in genetic network models. *Proceedings of the Second International Conference on Systems Biology.* Pasadena, California (2001) :222-230.
46. MJOLSNESS E, MANN T, CASTAÑO R, WOLD B: From coexpression to coregulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. *Advances in Neural Information Processing Systems*. (2000) 12, 928-934.
47. KOZA JR, MYDLOWEC W, LANZA G, YU J, KEANE MA: Reverse engineering of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing 2001.* World Scientific Publishing Co. (2001) 6, 434-445.
48. MAKI Y, TOMINAGA D, OKAMOTO M, WATANABE S, EGUCHI Y: Development of a system for the inference of large scale genetic networks. *Pacific Symposium on Biocomputing 2001.* World

Scientific Publishing Co. (2001) 6:446-458.
49. FRIEDMAN N, LINIAL M, NACHMAN I, PE'ER D: Using Bayesian networks to analyze expression data. *Proc. Fourth Annual Int. Conf. on Computational Molecular Biology*. ACM Press (2000) :127-135.
50. FRIEDMAN N, NACHMAN I, PE'ER D: Learning Bayesian network structure from massive datasets: The "Sparse Candidate" Algorithm. *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence* (1999) :206-215.
51. SPELLMAN P, SHERLOCK G, ZHANG M *et al.*: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* 9, 3273-3297 (1998).
52. FRIEDMAN N, LINIAL M, NACHMAN I, PE'ER D: Using Bayesian networks to analyze expression data. *J. Computational Biology* 7 (3/4), 601-620 (2000).
53. PE'ER P, REGEV A, ELIDAN G, FRIEDMAN N: Inferring subnetworks from perturbed expression profiles. *Bioinformatics.* Oxford University Press 1 (1), (2001).
54. BUTTE AJ, TAMAYO P, SLONIM D, GOLUB TR, KOHANE IS: Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* 97 (22), 12182-12186 (2000).
55. IDEKER TE, THORSSON V, KARP RM: Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing 2000.* World Scientific Publishing Co. 5, 302-313 (2000).
56. HARTEMINK AJ, GIFFORD DK, JAAKKOLA TS, YOUNG RA: Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing 2001.* World Scientific Publishing Co. 6, 422-433 (2001).
57. HARTEMINK AJ, GIFFORD DK, JAAKKOLA TS, YOUNG RA: Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing 2002.* World Scientific Publishing Co. 7, 437-449 (2002).
58. IMOTO S, GOTO T, MIYANO S: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pacific Symposium on Biocomputing 2002*. World Scientific Publishing Co. 7, 175-186 (2002).
59. DELAFUENTE A, BRAZHNIK P, MENDES P: A quantitative method for reverse engineering gene networks from microarray experiments using regulatory strengths. *Proceedings of the Second International Conference on Systems Biology.* Pasadena, California: 213-221 (2001).
60. YOO C, THORSSON V, COOPER GF: Discovery of causal relationships in a gene regulation pathway from a mixture of experimental and observational dna microarray data. *Pacific Symposium on Biocomputing 2002*. World Scientific Publishing Co. 7, 498-509 (2002).
61. IDEKER T, THORSSON V, RANISH JA *et al.*: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929-934 (2001).

• **Integrated approach to genetic network modeling.**

62. VON DASSOW G, MEIR E, MUNRO EM, ODELL GM: The segment polarity network is a robust developmental module. *Nature,* 406, 188-192 (2000).
63. MARNELLOS G, MJOLSNESS E: A gene network approach to modeling early neurogenesis in *Drosophila*. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 30-41, (1998).
64. KOSMAN D, REINITZ J, SHARP D: Automated assay of gene expression at cellular resolution. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 6-17 (1998).
65. KYODA K, KITANO H: Simulation of genetic interaction for *Drosophila* leg formation. *Pacific Symposium on Biocomputing '99.* World Scientific Publishing Co. 4, 77-89 (1999).
66. KYODA KM, MURAKI M, KITANOH: Construction of a generalized simulator for multi-cellular organisms and its application to smad signal transduction. *Pacific Symposium on Biocomputing 2000.* World Scientific Publishing Co. 5, 314-325 (2000).
67. MARNELLOS G, DEBLANDRE GA, MJOLSNESS E, KINTNER C: Delta-notch lateral inhibitory patterning in the emergence of ciliated cells in xenopus: experimental observations and a gene network model. *Pacific Symposium on Biocomputing 2000.* World Scientific Publishing Co. 5, 326-337 (2000).
68. MJOLSNESS E, SHARP D, REINITZ J: A connectionist model of development. *J. Theoretical Biology* 152 (4), 429-454 (1991).
69. KURHEKAR MP, ADAK S, JHUNHUNWALA S, RAGHUPATHYK: Genome-wide pathway analysis and visualization using gene expression data. *Pacific Symposium on Biocomputing 2002.* World Scientific Publishing Co. 7, 462-473 (2002).
70. PAVLIDIS P, LEWIS DP, NOBLE WS: Exploring gene expression data with class scores. *Pacific Symposium on Biocomputing 2002.* World Scientific Publishing Co. 7, 474-485 (2002).
71. STEPHENS M, PALAKAL M, MUKHOPADHYAY S, RAJE R, MOSTAFA J: Detecting gene relations from medline abstracts. *Pacific Symposium on Biocomputing 2001.* World Scientific Publishing Co. 6, 483-496 (2001).
72. WONG L: PIES, a protein interaction extraction system. *Pacific Symposium on Biocomputing 2001.* World Scientific Publishing Co. 6, 520-531 (2001).
73. BUSSEMAKER HJ, LI H, SIGGIA ED: Regulatory element detection using correlation with expression. *Nat. Genetics* 27, 167-171 (2001).
74. BRAZMA A, VILO J: Gene expression data analysis. *FEBS Letters* 480, 17-24 (2000).
75. TAVAZOIE S, HUGHES JD, CAMPBELL MJ, CHO RJ, CHURCH GM: Systematic determination of genetic network architecture. *Nat. Genetics* 22, 281-285 (1999).
76. VAN HELDEN J, ANDRE B, COLLADO-VIDES J: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Molecular Biology* 281, 827-842 (1998).
77. GUHATHAKURTA D, SCHRIEFER LA, HRESKO MC, WATERSTON RH, STORMO GD: Identifying muscle regulatory elements and genes in the nematode *Caenorhabditis Elegans*. *Pacific Symposium on Biocomputing 2002.* World Scientific Publishing Co. 7, 425-436 (2002).
78. DHAESELEER P, LIANG S, SOMOGYI R: Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16 (8), 707-726 (2000).

• **An early overview of clustering and genetic networks including data requirements.**

79. ALON U, BARKAI N, NOTTERMAN *et al.*: Broad patterns of gene expression revealed by cluster analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* (1999) 96:6745-6750.
80. BEN-DOR A, YAKHINI Z: Clustering

gene expression patterns. *J. Computational Biology* 6 (3/4), 281-297 (1999).

81. EISEN MB, SPELLMAN PT, BROWN PO, BOTTSTEIN D: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95 (25), 14863-14868 (1998).
82. MICHAELS GS, CARR DB, ASKENAZI M, FUHRMAN S, WEN X, SOMOGYI R: Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing '98.* World Scientific Publishing Co. 3, 42-53 (1998).
83. D'HAESELEER P, WEN X, FUHRMAN S, SOMOGYI R: Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. *Information Processing in Cells and Tissues*. Plenum Publishing: 203-212 (1998).
84. MURPHY K, MIAN S: Modeling gene expression data using dynamic bayesian networks. *Computer Science Division.* University of California (1999).
85. ARNONE A, DAVIDSON B: The hardwiring of development: organization and function of genomic regulatory Systems. *Development* 124, 1851-1864 (1997).
86. JEONG H, TOMBOR B, ALBERT R, OLTVAI ZN, BARABASI A-L: The large-scale organization of metabolic networks. *Nature* 407, 651-654 (2000).