Draft ISMB 01 Proc. Bioinformatics 17 Suppl 1 S270-S278 (2001)

BIOINFORMATICS **DISCOVERY NOTE**

Vol. 1 no. 1 2001
Pages 1–9

# *Computational Expansion of Genetic Networks*

Amos Tanay and Ron Shamir

School of Computer Science, Tel-Aviv University, , Tel-Aviv, 69978, Israel

**ABSTRACT**

We present a new methodology for computational analysis of gene and protein networks. The aim is to generate new educated hypotheses on gene functions and on the logic of the biological network circuitry, based on gene expression profiles. The framework supports the incorporation of biologically motivated network constraints and rules to improve specificity. Since current data is insufficient for de-novo reconstruction, the method receives as input a known pathway *core* and suggests likely *expansions* to it. Network modeling is combinatorial, yet data can be probabilistic. At the heart of the approach are a fitness function which estimates the quality of suggested network expansions given the core and the data, and a specificity measure of the expansions. The approach has been implemented in an interactive software tool called GENESYS. We report encouraging results in preliminary analysis of yeast ergosterol pathway based on transcription profiles. In particular, the analysis suggests a novel ergosterol transcription factor.

**Contact:** rshamir@tau.ac.il

## INTRODUCTION

Recent years witnessed an information revolution in biological research, following the advent of novel high throughput experimentation methods which encompass biological systems on a new scale. Most notable and mature of these methods is transcription profiling using oligonucleotide chips and cDNA micro-arrays, but other methods, like high throughput protein interactions, protein localizations and DNA binding assays, are developing rapidly. These methods create an urgent need for sophisticated computational tools that facilitate rapid and comprehensive analysis of large amounts of biological data.

Most of the computational analysis of micro-arrays experiments today is based on clustering of transcription profiles (Spellman et al., 1998; Eisen et al., 1998; Sharan and Shamir, 2000). Clustering is a useful way to identify common data patterns and its utility has been demonstrated in many studies. Still, it is a rather crude method, as it is based on pairwise comparisons, so clustering is only a first step of the data analysis. Deeper inference of relations at a higher level of complexity is called for, and is done mainly manually nowadays. The research on genetic networks is trying to shape a new methodology that will enable inference of more complex relations from the data.

Based on understanding of the biological regulatory mechanisms and on theoretical examination of the evolutionary implications of the system as a whole (Kauffman, 1993; Somogoyi and Sniegoski, 1996), researchers have constructed different mathematical models to describe the behavior of biological systems (Arkin et al., 1998; Chen et al., 1999; Dhaeseleer et al., 1999; Chen et al., 1999) (for a review see Dhaseleer et al. (2000)). Algorithms and complexity analysis of inferring a genetic network from experimental data were developed for some of these modeling approaches (Akutsu et al., 1998, 2000; Liang et al., 1998).

The discipline of genetic network analysis has not become yet a practical aid to the biologist. The major reason to this can be called "*experimental complexity*": Theoretical studies show that, without additional assumptions, the mathematical problem of inferring all but tiny genetic networks from experiments is impractical, since the number of experiments that would have to be performed in the worst case is out of reach (Akutsu et al., 1998). This is true even when the models assumed are simple Boolean networks. Although strong assumptions on the data reduce experimental complexity (e.g., random distribution in the attractor space, cf. Akutsu et al. (1999); Dhaseleer et al. (2000)), these assumptions do not hold for data gathered today. One still cannot expect enough data to support current reconstruction approaches in the foreseeable future.

The inherent (experimental) complexity of genetic network inference led researchers to create statistical tools that would reveal relevant biological features from available data (Friedman et al., 2000), and construct tools for an efficient design of an experiment plan to extract maximum information from a fixed laboratory "budget" (Karp et al., 1999; Ideker et al., 2000). From a different direction, Zein et al. (2000) suggested a method for comparative analysis in which a set of transcription profiles is analyzed against a set of known metabolic pathways, in order to identify which of them is manifest in the data. The last work was important in its utilization of known relations among genes, although the notion of metabolic pathway is not directly connected to regulatory function.

We present in this paper a novel framework for analysis of genetic networks and hypothesis generation. The

© Oxford University Press 2001

starting point of the process is a pathway core, which represents prior knowledge on a particular biological sub-system. A combinatorial search algorithm suggests the most promising core expansions, in light of their level of fitness to a given, heterogeneous experimental dataset. The use of a known core, together with the integration of data and additional biological constraints, reduce experimental complexity. They enable for the first time the computational generation of reasonable biological hypotheses, using datasets of realistic size that are already available today.

To support these ideas, we have formalized the notion of biological network models (generalizing Liang et al. (1998)) and adapted them to the representation of biological constraints and modern data sets. We have developed methods and algorithms to evaluate the fitness of a model vis-a-vis a set of given experiments, and studied the computational problem of finding an expansion of the core that would improve this fitness.

A new software platform, named GENESYS (GEnetic Network Expansion SYStem) was developed and used to test the framework with real biological information. We have focused on budding yeast and used publicly available transcription profile data to generate likely expansions of ergosterol related pathways. The results suggest a novel transcription factor and identify interesting regulation patterns, proving that computational analysis can reveal complex relations in genetic networks, even with today's data sets.

The paper is organized as follows. We first provide some definitions and notation to set up our modeling approach. Next we discuss algorithms for modeling fitness calculations and explore the computational problem of expanding a pathway core. Finally we present the results on real transcription profiles and pathways.

## MODELING: DEFINITIONS AND ASSUMPTIONS

In this section we describe the formal framework for our analysis and tools. We also explain the reasoning behind our modeling choices.

A **biological network** (or **model**) is defined by a set $U$ of **variables** (genes, proteins, mRNAs etc.), a set $C$ of **values** or **states** that the variables may attain, and **functions** $f^v : C^{\|U\|} \rightarrow C$ for each $v \in U$. The interpretation is that the value of variable $v$ at time $t$ depends on the values of its input variables at time $t-1$, and the functional dependence is described by $f^v$. We use the term **arguments** of $f^v$ for the non trivial arguments of the function. ($u$ is a trivial argument of $f_v$ if changing the value of $u$ alone never alters the value of the function.) We denote by $arg(f^v)$ the set of arguments of $f^v$.

A useful partial description of the network is given via its dependencies. The **dependency graph** of $N$ is a directed graph $G(N) = (U, A)$ where $(u, v) \in A$ iff $u$ is an argument of $f^v$). The set of arguments of $v$ in $G$ is $arg^G(v) = \{u|(u, v) \in A\}$.

Since we will be searching for a "best" network, we need to describe the search space next: A **model space** is defined by the four-tuple $(U, C, F_{bio}, G_{bio})$ where $U$ and $C$ are as above, $F_{bio} \subseteq \mathcal{F} := \{f : C^{\|U\|} \rightarrow C\}$ is the class of candidate functions, and $G_{bio}$ is a class of dependency graphs on $U$. The space consists of all networks with functions from $F_{bio}$ and dependency graphs from $G_{bio}$.

$F_{bio}$ and $G_{bio}$ are used to limit the model space, by incorporating biological knowledge and realistic constraints. $F_{bio}$ constrains the properties of each particular function. For example, $MONO_d$ is the set of monotone functions with at most $d$ inputs. $G_{bio}$ constrains the overall architecture of the network. For example, $INDEG_r$ is the set of graphs with indegrees at most $r$, and $MAXREG_r$ is the set of digraphs having at most $r$ nodes with outgoing edges (those nodes would be interpreted as the *regulators* of the network).

In our study of transcription regulation in yeast using gene expression data, the following model space was used: $U$ was the set of all mRNAs of yeast genes (ORFs). The values in $C$ corresponded to transcription changes: $-1$: down regulation; 0: normal; 1: up regulation. $G_{bio}$ was set to $INDEG_r$. There were no constraints on the candidate functions, i.e. $F_{bio} = \mathcal{F}$. See figure 1 for a concrete biological example of a network.

A key reason for distinguishing $G_{bio}$ and dependency graphs is that often we may have insufficient information to infer precise functional relations. Inferring dependencies only is less prone to over-fitting, yet it provides key information on the network.

In our definitions of a model space, values and time scale are discrete, and the functional relations are deterministic. This simplifying choice was made in order to reduce the number of degrees of freedom and to avoid over-fitting. We believe that important features of biological systems (mainly complex genetic and protein switches) can be elucidated using such simplified models. Continuous or stochastic modeling require rate constants tuning, which drastically increase the amount of information for validating a model. Such models are currently impractical to solve for all but very small networks.

For micro-array data, by an **experiment** we mean a triplet $(INP, OUT, PERT)$ where $INP$ and $OUT$ are the **input and output vectors**, assigning values to each variable in $U$. $PERT \subseteq U$ is the set of **perturbed variables**, i.e., those genes that were knocked-out or over-expressed. Hence, a knock-out or over-expressive experiment will produce one triplet. Time-series data, providing expression levels at a series of $n$ time points,

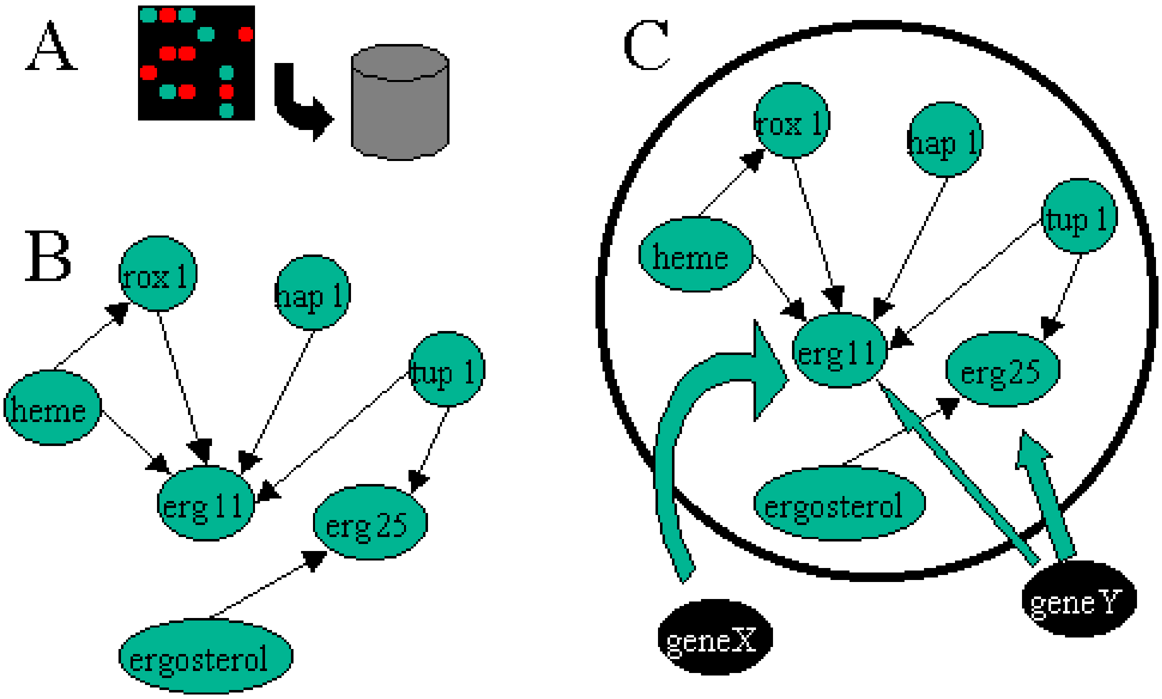


**Fig. 1.** Overview of the core expansion methodology. A) Large data sets (e.g. expression profiles) are transformed into a uniform database. B) A pathway core is constructed based on the literature. The core shown here is the dependency structure of part of the yeast ergosterol pathway, consisting of 7 variables. The exact logic of the system is defined by the association of a discrete function to each model variable. C) The core model is expanded with additional variables and interactions. Expansions are scored by their level of fitness to the database.

yield $n-1$ experiment triplets, where the vectors at time points $i$ and $i+1$ form $INP$ and $OUT$ of the $i$-th experiment. Note that this transformation assumes that data dependence is Markovian. We will use $INP^e_S$ ($OUT^e_S$) to denote the input (output) values of the variable set $S$ in the experiment $e$.

If in an experiment $INP = OUT$, we say it is a **steady state experiment**. Real data sets are often either time series of samples along some synchronized biological process (Spellman et al., 1998; DeRisi et al., 1997), or a single sample from a cell culture under some condition (Hughes et al., 2000). Steady state experiments might contain an averaging of an underlying temporal process and so modeling them correctly entails a less detailed representation of the biological system. Mathematically, for steady state data, one must exclude models with variables regulating themselves, in order to avoid the trivial self-regulation solution.

Some compensation for the discretization of the network space is provided by probabilistic modeling of the experimental data: (Below, and occasionally later, we use overlines on vectors for clarity.) A **(noisy) experiment** is a triplet $(\overline{PINP}, \overline{POUT}, PERT)$, where $PERT \subseteq U$, and $\overline{PINP}$ and $\overline{POUT}$ assign to each variable in $U$ a distribution over the values in $C$. In other words, $PINP_v(c)$ ($POUT_v(c)$) is the probability that $v$ attains the value $c \in C$ in the input (output). This enables better data utilization by factoring in the noise inherent in high throughput experiments.

## CORES, EXPANSION, AND FITNESS

In order to apply optimization strategies on the model space, one needs an objective function, which evaluates how well each network in the model space fits the experiments data. Often, we seek an optimal network that conforms with prior knowledge. To that end, we define a **core** as a network $N'$ defined on a subset $U' \subseteq U$. The core represents our prior knowledge. An **expansion** of a core is a network containing $N'$ as a subnetwork. Similarly, a **core digraph** is a dependency graph $G'$ defined on $U' \subseteq U$, a digraph containing $G'$ as a subgraph is an expansion of it.

Our goal is to infer biological pathways by finding an expansion network or digraph that fit the experimental data best. This must be preceded by developing a good fitness function. Such function should perform well both in sensitivity (scoring good expansions high) and specificity (scoring bad expansions low), and must also be efficiently computable. *Local* fitness functions evaluate the fit of the experimental data to the function $f^v$ of a single variable $v$, while *global* fitness evaluates the overall network. Our local fitness functions use ideas that generalize Liang et al. (1998):

Given a function $\phi \in F_{bio}$ and a set of experiments $E = (INP^e, OUT^e, PERT^e)_{e \leq n}$, the **consistency** of $\phi$ for variable $v$, or the consistency of the pair $(\phi, v)$, is:

$$Consist(v, \phi, E) = |\{e \in E, v \notin PERT_e \mid \phi(INP^e) = OUT_v^e\}|$$

Denote the arguments of $\phi$ by $x_1, \ldots, x_d$. The **consistency** of $(\phi, v)$ given a noisy experiment set $E$ is:

$$Consist(v, \phi, E) = Pr(\phi(x_1, \ldots, x_d) = v) = \sum_{e \in E} \sum_{u=(u_1,\ldots,u_k) \in C^k} (\prod_i PINP_{x_i}^e(u_i)) \times POUT_v^e(\phi(u))$$

The explicit formula assumes statistical independence of the distributions $\overline{PINP}_v$ and $\overline{POUT}_v$ for each $v$.

When seeking to infer dependencies only, we define $Consist(v, S, E)$, the consistency of a set $S$ of arguments for node $v$, as the maximum consistency obtained by any $f^v \in F_{bio}$ whose arguments all belong to $S$. An important special case is when there are no constraints on functions in the model space. In this situation we can compute the consistency of a candidate argument set for a node efficiently, as follows:

PROPOSITION 1. *For any $S = \{s_1, \ldots, s_d\} \in U$, if $F_{bio} = \mathcal{F}$ and $E$ is a set of noiseless experiments we have:*

$$Consist(v, S, E) = \sum_{c_1,\ldots,c_d \in C} max_{c \in C} |\{e \in E \mid INP_{s_i}^e = c_i, i = 1, \ldots, d \wedge OUT_v^e = c\}|$$

**Proof:** Since we have no constraints on the function once its set of arguments is determined, we can optimize the consistency by making the best choice for each input assignment independently. ■

A similar reasoning applies to noisy experiments, by maximizing the likelihood of the function value for each input assignment independently. Figure 2 outlines the algorithm for noisy experiments. The algorithm uses a table of size $|C|^{d+1}$ and iterates on the experiments to simultaneously sum the probabilities of all i/o transitions. Let $S = \{v_1, \ldots, v_d\}$, denote the number of experiments by $n$ and let $m = |C|$.

Consist($v$, $S$, $E$):
Initialize a $m^d \times m$ real valued table $vote$ and
a scalar $consist$ with zeroes.
**For** all $e \in E$ if $v \notin PERT^e$ do
  **For** all vectors $u \in C^d$ do
    $p_u = \prod_i INP_{v_i}^e(u_i)$
    For $i = 1, \ldots, m$ do:
      $vote[u, i] = vote[u, i] + p_u * POUT_v^e(c_i)$
**For** all $u \in C^d$ do:
  $consist = consist + max_i\{vote(u, i)\}$.

**Fig. 2.** Consistency computation.

PROPOSITION 2. *The consistency of an argument set for a variable in the model space $(U, C, \mathcal{F}, INDEG_d)$ can be computed in $O(nm^d(m + d))$ steps for noisy experiments and $O(n+m^{d+1})$ steps for perfect experiments.*

Though simple and easy to compute, the consistency function gives no information regarding the specificity of a speculated regulation pattern, and is thus very sensitive to over fitting. To address this problem, we shall describe how to calculate a "p-value" of the measured consistency $\kappa$ = Consist($v$, $S$, $E$). As our null hypothesis, we assume independence of the measured values of the variable $v$ and the variables in $arg(v)$. We wish to estimate the probability of observing consistency $\kappa$ or higher in the data under the null hypothesis. Consider first the case of perfect experiments and assume $v$ was not perturbed in any experiment. Now define a probability space based on the data. We use two random variables, $X$ and $Y$. $X$ attains values in $C$ with probabilities :

$$p_i = Pr(X = c_i) = \frac{1}{n} |\{e \in E \mid OUT_v^e = c_i\}| \quad (1)$$

If $arg(v) = \{v_1, \ldots, v_d\}$, $Y$ is taking values in $C^d$ with probabilities :

$$Pr(Y = (c_1, .., c_d)) = \frac{1}{n} |\{e \in E | INP_{v_i}^e = c_i, i = 1, \ldots, d\}| \quad (2)$$

Let $S$ be a set of possible arguments for a variable $v$, with consistency $\kappa$. The **regulation specificity** of the pair $(S, v)$, denoted $rSpec(S, v, E)$, is the probability of obtaining a consistency of $\kappa$ or higher in the probability space $(Y \times X)^n$. Note that one can also use $rSpec$ itself as a fitness function (with a negative sign, to maintain the formulation of maximizing fitness).

The size of the probability space defined above, is exponential in $n$, so a naive algorithm for computing $rSpec$ is not practical. We present an approximation which is practical when $n - \kappa$ is small (almost perfect consistency) and is linear in the number of experiments $n$. We use a random variable from the space $X$ defined above and set the input values deterministically to $INP_S$. We now calculate the probability $\pi(\kappa)$ for obtaining a consistency $\kappa$ or better in a data set with the $n$ inputs from $INP_S$ and outputs sampled from $X$. If there are $l$ input configurations with multiplicities $m_1, \ldots, m_l$ then $\pi(\kappa)$ is the probability of getting $n'_i$ identical values out of $n_i$ samples from $X$ for $i = 1, \ldots, l$ and $\sum n'_i \geq \kappa$.

Denote by $\psi(r, s)$ the probability of getting at least $s$ identical values when sampling $r$ times from $X$. Then $\psi(r, s)$ can be computed exhaustively in $O(rm^r)$ time. To compute $\psi(r, s)$ efficiently we distinguish two cases:

(1) if $s \geq \frac{r}{2}$ then $\psi(r, s) = \sum_j \sum_{i=0}^{s} \binom{s}{i} p_j^{s-i}(1 - p_j)^i$ so is computable in $O(mr)$ time.

(2) if $s \leq \frac{r}{2}$ then $r \geq 2s$ or $2r - 2s \geq r$, so $rm^r = O(rm^{2(r-s)})$.

So in particular, when $r - s = O(1)$ computing $\psi(r, s)$ is polynomial in $r$ and $m$ for all $s$.

To compute $\pi(\kappa)$ we enumerate all integer partitions of $t = n - \kappa, t_1, \ldots, t_l$ s.t. $\sum t_i = t$, and compute:

$$\pi(\kappa) = \sum_{(t_i)} \prod_i \psi(n_i, n_i - t_i) \tag{3}$$

Since $n_i - t_i \leq n - \kappa$ we get that $\pi(k)$ is computable in $O(t^l(m^{2t} + l))$ time, so we conclude

PROPOSITION 3. *If $t = n - \kappa = O(1)$ then the regulation specificity of $(S, v)$ is computable in $O(n + t^l(m^{2t} + l))$ time.*

The generalization of p-values to noisy experiments is done as follows. Given a noisy experiments set $\overline{E} = \{e_i\}_{i<n}$ we assume statistical independence of the distributions $\overline{PINP}$ and $\overline{POUT}$ and construct a probability space that represents possible deterministic instantiations of the noisy experiment set. We define a random variable $E_r$ with values in the space of perfect experiments sets. The probability of obtaining a given perfect experiment set value $\{h_i\}_{i<n}$ is:

$$Pr(\{h_i\}) = \prod_{v \in U, i<n} PINP_v^{e_i}(INP_v^{h_i}))POUT_v^{e_i}(OUT_v^{h_i}))$$

Then

$$rSpec(v, G, E) = E(rSpec(v, G, E_r)) \tag{4}$$

Again, a naive computational approach for the finding the expectations above is impractical. We have performed approximate evaluations by exhausting only part of the probability space for $E_r$.

We are now ready to state our main optimization problem: The **pathway expansion problem** is defined with respect to a model space $(U, C, F_{bio}, G_{bio})$ and using a prescribed fitness function $fit$. Given a set of experiments $E$ and a core digraph $G' = (U', E')$, find a core expansion $G'' \supseteq G'$ maximizing $fit(G'')$. If several solutions exists, find one minimizing $\|G''\|$.

For an expansion $G''$ set $S_v = \{x \in G'' | xv \in E''\}$ and define

$$fit(G'') = \sum_{v \in U'} consist(v, S_v, E) \tag{5}$$

PROPOSITION 4. *The pathway expansion problem, with the fitness function (5), is NP hard, even assuming constant time computation of fitness, and even for cores of size one.*

**Proof:** We shall show that the decision version of the problem, "is there an expansion with perfect consistency and size $\leq l$ ?" is NP-complete. Clearly that problem is in NP. We will construct a reduction from SET COVER. Given a set $S = \{a_1, \ldots a_r\}$ and a collection of subsets $I = \{S_1, \ldots, S_q\}$ of $S$. Construct an instance of the expansion problem as follows. $U$ will be the set of subsets plus an additional variable, i.e. $U = \{1, \ldots, q, "c"\}$. The experiments set will consists of $n + 1$ steady state experiments indexed by $S \cup "0"$ and defined by the matrix below (columns are variables, rows are experiments, $\chi$ is the standard subset characteristic function):

$$\begin{array}{c|cccc} & c & 1 & .. & q \\ \hline 0 & 0 & 0 & .. & 0 \\ 1 & 1 & \chi_{S_1}(a_1) & \ldots & \chi_{S_q}(a_1) \\ \vdots & \vdots & & & \\ r & 1 & \chi_{S_1}(a_r) & \ldots & \chi_{S_q}(a_r) \end{array} \tag{6}$$

The core is set simply to the single variable $c$ and we set $l = k + 1$.

We will show that an expansion of $c$ with perfect $c$ consistency is equivalent to a set cover. First note that any set of expansion variables is equivalent to a collection of subsets $I' \subset I$. Now if $c$ is perfectly consistent with $I'$ then there do not exist $e_1, e_2$ such that $\overline{/INP_{I'}^{e_1}} = \overline{INP_{I'}^{e_2}}$ and $OUT_c^{e_1} \neq OUT_c^{e_2}$. Taking $e_2$ as the "0" experiment implies that there is no $e$ s.t. $\overline{INP_{I'}^{e}} = \overline{INP_{I'}^{0}}$. Since $\overline{INP_{I'}^{0}}$ is a vector of zeros, we conclude that for each $e \in E - 0$ (equivalent to an $S$ element) we must have a variable in the expansion $I'$ (equivalent to a subset in the cover) with non zero value (equivalent to having a subset covering the element).

Now assume I' is a set cover, taking the set as an expansion yield perfect consistency since the only experiment with 0 values over all the expansion is the ”0” experiment (otherwise the node represented by the experiment is not covered).

In conclusion, there exist a set cover $I'$ with $\|I'\| \leq k$ iff there exist an expansion $U'' = I' \cup \{c\}$ s.t. $\|U''\| \leq k+1$. ■

In the case of bounded indegree, the pathway expansion problem is polynomial: If all indegrees are at most $d$, then the set of candidates for expansion is of size at most $\|U'\| * d$, and trying all such sets is polynomial for fixed $d$.

## RESULTS

GENESYS (GEnetic Network Expansion SYStem) is a new software platform implementing the concepts and methods described above. The environment includes engines for representing networks and computing fitness, a flexible expansion algorithm, viewers for visualization of biological data sets, an application to enable interactive usage of the viewers and engine and an internal database scheme for the storage of datasets and pathways.

The system was implemented in C++ and Perl/Tk under linux (about 25000 code lines). It is able to analyze single node expansions (see below) of cores with up to 30 nodes within ten minutes or less on a standard pc.

To test our ideas, we applied GENESYS to yeast transcription datasets using the ergosterol pathway as a core. We focused on the simplest possible core expansion: The **single node expansion** process examines each of the variables in $U$ and calculates the sum of fitness gains to all core variables from adding that variable to the core. Note that unlike clustering or similarity tests, we are not looking for genes that are similar across the entire data set, but rather seek genes that might regulate or indirectly affect the pathway in those experiments which are left unexplained by the core model. We present below the results of two different screening processes, with different limitations and goals.

The fitness function was computed as follows. Denote the core by $U'$. For each non-core variable $v$, its global fitness is

$$\sum_{u \in U'} \max_{S \subset U \cup \{v\}, |S| \leq d} -rSpec(u, S, E).$$

We have actually used a variant of $rSpec$ that is more efficiently computable, and yields essentially the same results as reported below.

### Preprocessing Expression Data

Our study focused on yeast, which has the largest publicly available gene expression datasets. The variable set consisted of 6200 yeast ORFs. Transcription profiles were taken from two large scale yeast cDNA arrays experiments: Hughes et al. (2000) performed some 260 selected knock-out experiments ; Gasch et al. (2000) performed 100 experiments testing yeast behavior in stressful conditions. We have chosen to view all experiments as steady state experiments: For knockout experiments this is a natural choice. For the stress time series data, we chose to view each measured transcription profile as a different steady state experiment, since the time intervals between measurements were non uniform and typically much larger than the transcription activation delay.

Details on data normalization and transformation into probability distributions can be found in the supplementary web page (www.math.tau.ac.il/˜rshamir/genesys/ismb01).

Pathway cores were generated based on available literature, notably SGD (Ball et al., 2001) and YPD (Costanzo et al., 2001). Other biological references will be cited when discussing the results.

### Ergosterol Metabolism

Ergosterol is an essential lipid in yeast which is similar to cholesterol in mammals. Ergosterol’s primary role is in the cell membranes but it is also involved in aerobic metabolism, sterol uptake and sterol transport. Ergosterol metabolism is understood rather well. As many of the knockout experiments of Hughes et al. (2000) targeted that pathway, and it is believed to undergo significant transcription regulation, we chose it to test our analysis techniques. Ergosterol metabolism is composed of two pathways in series. The first, the mevalonate pathway, transforms acetyl-CoA to farnesyl and provides essential components for few important metabolic pathways (e.g. heme and quinones). The latter part transforms farnesyl to ergosterol. Much of the regulation of ergosterol is believed to be transcriptionally mediated, but the actual details are known only in part (Daum et al., 1998; Bammert and Fostel, 2000; Turi and Loper, 1992).

Figure 3 shows the basic known ergosterol metabolic pathway from farnesyl to ergosterol, including a series of 11 enzymes and three transcription factors. It is important to stress here the difference between metabolic pathways and regulatory networks: The fact that two enzymes follow each other in a biochemical process does not mean their *transcription* regulation is directly connected. We have modeled the ergosterol dependency structure core as the set of variables, with dependencies marked only between known transcription factors and their targets. In other words, no dependency was prescribed between enzymes. We have used this core and the expression data described above to test GENESYS.

We have analyzed the reactions of pathway enzymes in the entire data set (see supplementary data). A number of experiments showed a global reaction of the pathway:

in those experiments most of the pathway enzymes underwent significant change. This is presumably the result of some self regulatory mechanism (and indeed ergosterol itself is reported to function as transcription regulator for its pathway enzymes). However, many other experiments (about 40) showed a change in one or more of the pathway genes, which is not explained by the above mechanism. Those experiments may be explained by a more elaborate model. This motivates our attempt to expand the model and explain more of the data.

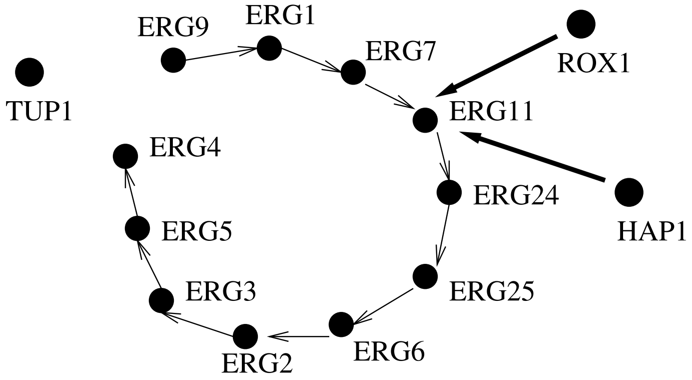


**Fig. 3.** The ergosterol pathway from farnesyl to ergosterol. Only enzymes (names starting with ERG) and known transcription factors (ROX1, HAP1, TUP1) are shown. Thin arrows indicate subsequent enzymes (not a model dependency). Thick arrows indicate model dependencies

## Transcription Factors Screening

Out of the ˜6200 yeast ORFs, we identified 130 putative transcription factors (TFs). For this we used SGD annotations, as well as typical structural motifs (e.g., zinc fingers). We then applied the single node expansion algorithm, limiting the candidates for node expansions to these putative TFs. In the first test, we ranked the fitness gain of each of the putative TFs against a ”naked” core consisting of the eleven ERG enzymes with no dependencies among them. HAP1 was ranked second out of 130 (Table 1), in agreement with the known role of HAP1 in ERG11 regulation. TUP1 is a general repressor and was thus ranked lower, ROX1 was less expressed in the data and was ranked much lower.

Having gained some confidence in the process quality, we focused on improving our understanding of ERG11 regulation. Turi and Loper (1992) analyzed the promoter region of ERG11 with results that are summarized in Figure 4. This time we applied the single node expansion to a core consisting of the eleven ERG enzymes as well as HAP1 and ROX1 as regulators of ERG11. The algorithm

| Gene | Annotation | Gain |
|---|---|---|
| PIP2 | Peroxisome proliferation | 0.0866 |
| HAP1 | erg11 activator | 0.07 |
| YDR213W | Unknown | 0.0624 |
| GLN3 | nitrogen catabolite | 0.0602 |
| RAP1 | transcription | 0.0547 |

**Table 1.** Putative transcription factors that ranked best in an expansion of the ”naked” ergosterol core

measured the improvement in fitness contributed by each of the 130 TFs, and an uncharacterized gene was ranked first. That gene improves the fitness of ERG11 (and others). Remarkably, it also has a good homology to HAP1 (33% identity, 50% similarity along 100 amino acids and even better in a shorter range). Moreover, analyzing ERG11 logic as a function of HAP1, ROX1, TUP1 and the novel TF shows that the effect of the new putative TF on ERG11 is inductive (as expected from a UAS2 binding gene). We thus have evidence from three different methods: sequence homology, promoter analysis indicating a second inducer should exist, and our screening procedure using some 360 different expression profiles in distinct cell states. All three support the hypothesis that our novel TF is indeed an ERG11 regulator that might bind to UAS2. We are in the process of testing this hypothesis experimentally.

## Screening All Genes

The admission of putative transcription factors only as added variables was important in the reduction of model space, and it allowed us to obtain very specific results. It is, however, interesting to try and screen all ˜6200 yeast ORFs against the ergosterol core. This type of analysis may discover more general patterns of regulation that cannot be directly tagged as ”A is a factor of B”. Still, as shown below, some interesting biology may be learned from it. The results of such a screen are given in Table 2. The two top ranking genes, POS5, YBR043C, are both of unknown function. POS5 has homology to iron metabolism enzymes. Both present significant fitness gain for ERG4 regulation. ERG4 is the last of the ergosterol pathway enzymes, is not essential and little is known on its regulation. Figure 5 gives a more detailed look on the relations among the three genes. Note that using standard clustering or similarity, the behavior of ERG4 in experiments with no POS5, YBR043C involvement would have masked the pattern identified here.

The fourth gene in the screening list is INO1 which is involved in inositol biogenesis. Inositol has a regulatory

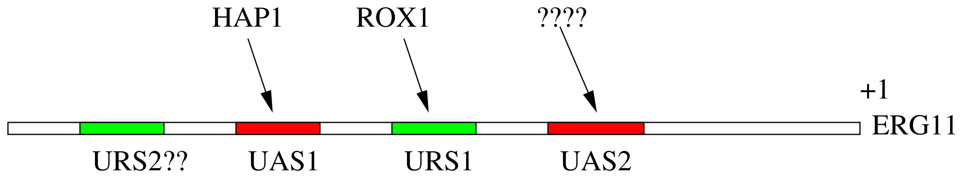


**Fig. 4.** ERG11 promoter region according to (Turi and Loper, 1992). UAS/URS: upstream activation/repression site. The transcription factors HAP1 and ROX1 induce and repress, respectively, ERG11 transcription via the binding sites UAS1 and URS1. UAS2 was identified as a likely binding site of an unknown activator. One of our goals in this study was to demonstrate that we can suggest the identity of the missing activator.

| Gene | ORF | Annotation | Gain | Gain location |
|---|---|---|---|---|
| 1.POS5 | YPL188W | Unknown | 0.0260306 | ERG4 |
| 2.YBR043C | YBR043C | Unknown | 0.0234197 | ERG4 |
| 4.INO1 | YJL153C | Inositol biosynthesis | 0.018337 | ERG6,ERG25,ERG5 |
| 7.GAS1 | YMR307W | cell surface glycoprotein | 0.016935 | ERG4,ERG6,ERG5,ERG25 |
| 10.MKK2 | YPL140C | PCK1 signaling | 0.016416 | ERG4 |
| 11.ERG10 | YPL028W | Ergosterol metabolism | 0.0162954 | ERG6,ERG25,ERG5,ERG7 |

**Table 2.** Results of 1-expansion of the ergosterol core pathway. Gene annotations are from SGD. 'Gain': the increase to fitness by using the additional variable. 'Gain location': the core genes whose regulation modeling was significantly improved by the variable, in order of significance.

function in the phospholipid pathway (adjacent to ergosterol). Note that the dependency is localized differently (improving different variables) in that case. The relation of GAS1 to ergosterol might be rooted in its function in the cell wall. The dependency between our core and MKK2 is very reasonable considering its function in the signaling pathway to the cell wall protein PCK1. The 11th gene in the list is ERG10, which is the first gene in the mevalonate pathway leading to our core.

The dependencies revealed by the general 1-expansion screening can serve as the basis for deeper biological exploration. The process pinpoints statistically significant patterns which are hard to identify otherwise. In contrast with the TF 1-expansion screening, the results are less direct and do not identify specific dependencies.

## CONCLUDING REMARKS

We have presented a new methodology for biological hypotheses generation, using genetic network cores and high throughput experimental data. A new software platform, called GENESYS, was implemented to enable analysis of available transcription profiles data sets and target pathways. Several initial test cases with the ergosterol pathway and yeast transcription profiles show that correct hypotheses are generated. We were able to find several biologically interesting regulation patterns including a novel putative ergosterol transcription factor.

GENESYS is under continuous development. Future goals would be to further improve global fitness calculation, to test the system on additional pathways and data sets, and to prepare the theory and tools for the incorporation of large scale Proteomics data.

## ACKNOWLEDGMENTS

We thank Martin Kupiec, Itsik Pe'er and Zohar Yakhini for helpful discussions. This research was supported in part by a grant from Agilent Technologies.

## REFERENCES

Akutsu, T., S. Kuhara, O. Maruyama, and S. Miyano (1998). Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 695–702.

Akutsu, T., S. Miyano, and S. Kuhara (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Proceedings of the 1999 Pacific Symposioum in Biocomputing (PSB 99)*, pp. 17–28.

Akutsu, T., S. Myano, and S. Kuhara (2000). Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint functions. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 00)*, pp. 8–14.

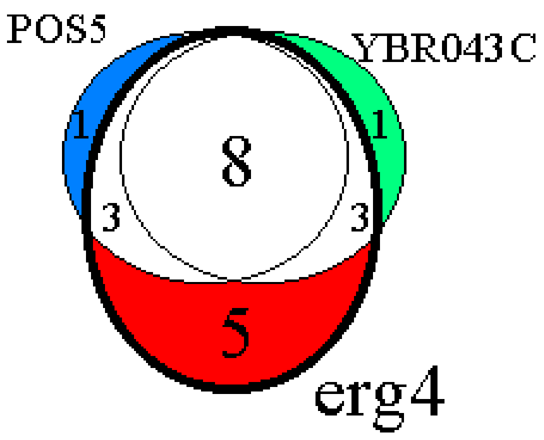


**Fig. 5.** ERG4 dependent genes. The Venn-like diagram represents all experiments in which ERG4,POS5 and YBR043C were induced. The number inside each of the sets indicate its size. Induction in this case is any up-regulation with p-value less than 0.01. The graph shows that induction of POS5 and YBR043C strongly correlate with ERG4 induction (11/12 experiments in both cases). ERG4 is showing a second, separate regulation pattern (5 experiments) which is unrelated to POS5,YBR043C

Arkin, A., J. Ross, and H. McAdams (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda infected escherichia coli cells. *Genetics 149*, 1275–1279.

Ball, C. et al. (2001). Saccharomyces genome database provides tools to survey gene expression and functional analysis data. *Nucleic Acids Research 29*, 80–1.

Bammert, G. and J. Fostel (2000). Genome-wide expression patterns in Saccharomyces cerevisiae: comparison of drug treatments and genetic alterations affecting biosynthesis of ergosterol. *Antimicrobial Agents and Chemotherapy 44*, 1255–1265.

Chen, T., V. Filkov, and S. S. Skiena (1999). Identifying gene regulatory networks from experimental data. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pp. 94–103.

Chen, T., H. L. He, and G. M. Church (1999). Modeling gene expression with differential equations. In *Proceedings of the 1999 Pacific Symposioum in Biocomputing (PSB 99)*, pp. 29–40.

Costanzo, M. et al. (2001). YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Research 29*, 75–9.

Daum, G., N. Lees, M. Bard, and R. Dickson (1998). Biochemistry, cell biology and molecular biology of lipids of Saccharomyces cerevisiae. *Yeast 14*, 1471–1510.

DeRisi, J., V. Iyer, and P. Brown (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science 278*, 680–686.

Dhaeseleer, P., X. Wen, S. Fuhrman, and R. Somogoyi (1999). Linear modeling of mRNA expression levels during CNS development and injury. In *Proceedings of the 1999 Pacific Symposioum in Biocomputing (PSB 99)*, pp. 41–52.

Dhaseleer, P., S. Liang, and R. Somogoyi (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics 16*, 707–726.

Eisen, M., P. Spellman, P. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA 95*, 14863–14868.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 00)*, pp. 127–135.

Hughes, T. et al. (2000). Functional discovery via a compendium of expression profiles. *Cell 102*, 109–26.

Ideker, T., V. Thorsson, and R. Karp (2000). Discovery of regulatory interaction through perturbation: inference and experimental design. In *Proceedings of the 2000 Pacific Symposioum in Biocomputing (PSB 00)*, pp. 305–316.

Karp, R. M., R. Stoughton, and K. Y. Yeung (1999). Algorithms for choosing differential gene expression experiments. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pp. 208–217.

Kauffman, S. (1993). *The Origins Of Order, Self Organization and Selection in Evoultion*. Oxford University Press.

Liang, S., S. Fuhrman, and R. Somogoyi (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of the 1998 Pacific Symposioum in Biocomputing (PSB 98)*, pp. 18–29.

Sharan, R. and R. Shamir (2000). CLICK: a clustering algorithm with applications to gene expression analysis. In *Proceedings of the Eighth Annual Conference on Intelligent Systems for Molecular Biology (ISMB 00)*, pp. 307–316.

Somogoyi, R. and C. Sniegoski (1996). Modeling the complexity of genetic networks. understanding multigene and pleiotropic regulation. *Complexity 1*, 45–50.

Spellman, P. et al. (1998). Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell 9*, 3273–3297.

Turi, T. and J. Loper (1992). Multiple regulatory elements control expression of the gene encoding the Saccharomuces cerevisiae Cytochromoe P450, lanosterol 14a-demethylase (ERG11). *JBC 267*, 2046–56.

Zien, A., R. Kuffner, R. Zimmer, and T. Lengauer (2000). Analysis of gene expression data with pathway scores. In *Proceedings of the Eighth Annual Conference on Intelligent Systems for Molecular Biology (ISMB 00)*, pp. 407–417.