BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–9

# Functional Annotation of Regulatory Pathways

Jayesh Pandey[a*], Mehmet Koyutürk[a], Yohan Kim[b], Wojciech Szpankowski[a], Shankar Subramaniam[b], Ananth Grama[a]

[a] Department of Computer Science, Purdue University
[b] Department of Chemistry and Biochemistry, University of California, San Diego.

**ABSTRACT**

**Motivation:** Standardized annotations of biomolecules in interaction networks (*e.g.*, Gene Ontology) provide comprehensive understanding of the function of individual molecules. Extending such annotations to pathways is a critical component of functional characterization of cellular signaling at the systems level.

**Results:** We propose a framework for projecting gene regulatory networks onto the space of functional attributes using multigraph models, with the objective of deriving statistically significant pathway annotations. We first demonstrate that annotations of pairwise interactions do not generalize to indirect relationships between processes. Motivated by this result, we formalize the problem of identifying statistically over-represented pathways of functional attributes. We establish the hardness of this problem by demonstrating the non-monotonicity of common statistical significance measures. We propose a statistical model that emphasizes the modularity of a pathway, evaluating its significance based on the coupling of its building blocks. We complement the statistical model by an efficient algorithm and software, NARADA, for computing significant pathways in large regulatory networks. Comprehensive results from our methods applied to the *E. coli* transcription network demonstrate that our approach is effective in identifying known, as well as novel biological pathway annotations.

**Availability:** NARADA is implemented in Java and is available at `http://www.cs.purdue.edu/homes/jpandey/narada/`.

**Contact:** Jayesh Pandey, `jpandey@cs.purdue.edu`.

## INTRODUCTION

Gene regulatory networks represent powerful formalisms for modeling cell signaling. These networks are inferred from gene expression, as well as other sources of data, using various statistical and computational methods (Friedman *et al.*, 2000; Husmeier, 2003). Recent studies on networks of specific organisms show that interactions between genes that take part in specific pairs of biological processes are significantly overrepresented (Lee *et al.*, 2002; Tong *et al.*, 2004). Generalizing such observations to pathways of arbitrary length may allow identification of standardized pathways, enabling creation of reference databases of direct and indirect interactions between various processes. Knowledge of such pathways is useful, not only in general understanding of the relationship between cellular processes at the systems level, but also in projecting existing knowledge of cellular organization of model organisms to other species. Increasing availability of species-specific interaction data, coupled with attempts aimed at creating standardized dictionaries of functional annotation for biomolecules provide the knowledge base that can be effectively used for this purpose. What is lacking is a comprehensive set of tools that combine these two sources of data to identify significantly over-represented patterns of interaction through reliable statistical modeling with a formal computational basis.

In this paper, we introduce the notion of *functional network characterization*, derived from a gene regulatory network and associated functional annotations of genes. We use the Gene Ontology (GO) (Ashburner *et al.*, 2000) for annotations, however, our methods themselves generalize to other networks and annotations. Functional network characterization is based on the *abstract* notion of regulatory interactions between pairs of functional attributes (as opposed to genes). In this context, we demonstrate that methods for identifying significant pairwise annotations do not generalize to pathway annotations. We introduce the problem of identifying statistically over-represented *pathways* of functional attributes, targeted at the identification of chains of regulatory interactions between functional attributes. We study the hardness of this problem, focusing on the non-monotonicity of commonly used statistical significance measures. We show that the problem is hard along two dimensions: (i) the pathway space of the functional attribute network, and (ii) the space of functional resolution specified by GO hierarchy. Emphasizing the modularity of a pathway to assess its significance, we propose a statistical model that focuses on the coupling of the building blocks of a pathway. We use this statistical model to derive efficient algorithms for solving the pathway annotation problem. Our methods are implemented in a web-based tool, NARADA which provides an intuitive user and data interface.

Comprehensive evaluation of NARADA on an *E. coli* transcription network from RegulonDB (Salgado *et al.*, 2006) shows that our method identifies several known, as well as novel pathways, at near-interactive query rates. Note that the current knowledge of regulatory networks is incomplete, and limited to a few model organisms. Therefore, the application of our method on currently available data does provide a comprehensive library of regulatory network annotation. On the other hand, the partial annotation provided by our method forms a useful basis for extending our knowledge of regulatory networks beyond well-studied processes and model organisms.

## BACKGROUND AND MOTIVATION

*Results from previous studies.* Lee *et al.* (2002) study the *S. cerevisiae* transcription regulation network with a view to understanding relationships between functional categories. They observe that transcriptional regulators within a functional category commonly bind to genes encoding regulators within the same category (*e.g.*, cell

*to whom correspondence should be addressed

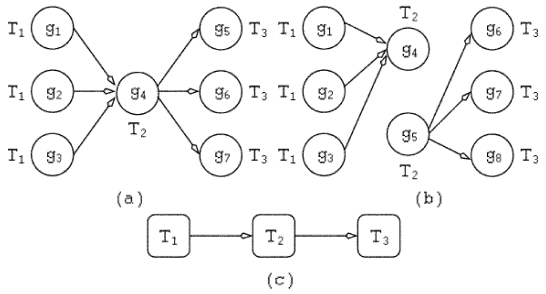


**Fig. 1.** Pairwise assessment of regulatory interactions between functional attributes may result in identification of non-existent patterns. Two regulatory networks are shown in (a) and (b). The nodes are labeled by corresponding genes and each gene is tagged with a set of functional attributes. The network of functional attributes resulting from both networks, considering only pairwise interactions, is shown in (c).

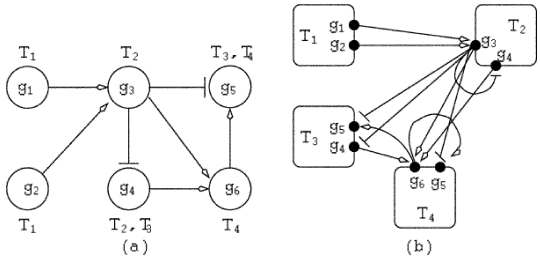


**Fig. 2.** (a) A sample gene regulatory network and the functional annotation of the genes in this network. Each node represents a unique gene and is tagged by the set of functional attributes attached to that gene. Activator interactions are shown by regular arrows, repressor interactions are shown by dashed arrows. (b) Functional attribute network derived from the gene regulatory network in (a). In this multigraph, nodes (functional attributes) are represented by squares and ports (genes) are represented by dark circles.

cycle, metabolism, environmental response). They also report that many transcriptional regulators within a functional category bind to transcriptional regulators that play key roles in the control of other cellular processes. For example, cell cycle activators are observed to bind to genes that are responsible for regulation of metabolism, environmental response, development, and protein biosynthesis. Tong *et al.* (2004) identify putative genetic interactions in yeast via synthetic genetic array (SGA) analysis and investigate the functional relevance of their results in the context of GO annotations. They construct a network of GO terms by inserting an edge between any pair of terms that are *bridged* by a significant number of interacting gene pairs. Here, two GO terms are said to be bridged by an interaction if one of the interacting genes is associated with one of the terms, and the other gene with the second term, but neither is associated with both terms. They show that the resulting network is clustered according to underlying biological processes, while some biological processes buffer one another. For example, microtubule-based functions buffer both actin-based and DNA synthesis or repair functions, suggesting coordination of these functions through interaction of various genes.

*Approach.* Establishing functional relationships from gene interactions is essential to understanding functional organization of a cell. Current investigations are limited to case-specific studies that generally focus on validation or evaluation of results through simple statistical analyses – yet they provide significant insights (Lee *et al.*, 2002; Tong *et al.*, 2004; Gamalielsson *et al.*, 2006). Computational tools that are based on sophisticated abstractions and customized statistical models are likely to yield novel insights. The basic approach for integrating existing knowledge of gene networks and functional annotations is to project the network in the *gene space* onto the *functional attribute space* through mapping of genes to attributes as specified by the annotation. A simple method for achieving this annotates each gene with its function and identifies overrepresented interacting annotations. This method yields interesting insights, as illustrated by Tong *et al.* (2004) in the context of synthetic genetic arrays. This model, however, does not generalize beyond pairwise interactions since each interaction between a pair of functional attributes is within a specific context (a different pair of genes) in the network, as illustrated by the following example.

*Motivating example.* Two regulatory pathways are shown in Figure 1 – each node is identified by its corresponding gene ($g_i$) and tagged by the functional attribute ($T_j$) associated with the gene. In Figure 1(a), genes $g_1$, $g_2$, and $g_3$ indirectly regulate genes $g_5$, $g_6$, and $g_7$ through gene $g_4$. In Figure 1(b), the network is isolated and there is no indirect regulation. Now assume the network of functional attributes derived from the simple method described above, separately for each gene network. For both networks, since all genes associated with functional attribute $T_1$ regulate a gene with $T_2$, one may conclude that $T_1$ regulating $T_2$ is significant. A similar conclusion follows for the regulatory effect of $T_2$ on $T_3$. If only pairwise interactions are considered, we derive the same network of functional attributes from both genetic networks (Figure 1(c)). This network clearly suggests that functional attribute $T_1$ indirectly regulates $T_3$ through $T_2$. This is indeed a correct observation for the network in Figure 1(a). However, this is not true for the network in Figure 1(b).

To address this problem, we develop a formal framework for projecting a gene network on to a network of functional attributes, using *multigraph* models that accurately capture the context in which an interaction occurs. Through this framework, we generalize pairwise interactions between functional attributes to the identification of regulatory pathways of functional attributes.

## METHODS

We now describe the biological, statistical, and computational formalisms that underly our methods.

### Formal Model for Functional Attribute Networks

A *gene regulatory network* is modeled by a labeled directed graph $G(V_G, E_G, M_G)$. In this network, nodes $g_i \in V_G$ represent genes. Directed edge $g_i g_j \in E_G$, where $g_i, g_j \in V_G$, represents a regulatory interaction between genes $g_i$ and $g_j$. $M_G : E_G \rightarrow \{+, -, \pm\}$ specifies a labeling of edges that represents the mode of regulation: activation ($+$), repression ($-$), or dual regulation ($\pm$). A sample gene regulatory network is shown in Figure 2. In our discussion, for the sake of simplicity, we omit the mode of regulation and treat all interactions as activator interactions, whenever appropriate.

Each gene in the network is associated with a set of *functional attributes*. These attributes describe a functional *annotation* of

the gene, *i.e.*, they map an individual biological entity to known functional classes.

DEFINITION 1. **Functional Annotation.** *Given a set of genes $V_G$ and a set of functional attributes $V_F$, let $2^{V_G}$ and $2^{V_F}$ denote the power set of $V_G$ and $V_F$, respectively. Then, functional annotation $\mathcal{A}(V_G, V_F) = \{\mathcal{F}, \mathcal{G}\}$ defines mapping $\mathcal{F} : V_G \rightarrow 2^{V_F}$ and $\mathcal{G} : V_F \rightarrow 2^{V_G}$, such that $T_j \in \mathcal{F}(g_i)$ if and only if $g_i \in \mathcal{G}(T_j)$, for any $g_i \in V_G$ and $T_j \in V_F$. The frequency of $T_j$, $\phi(T_j) = |\mathcal{G}(T_j)|$, is equal to the number of genes that are mapped to $T_j$.*

In Figure 2(a), each gene $g_i$ is tagged with the functional attributes in $\mathcal{F}(g_i)$. For each $T_j$, $\mathcal{G}(T_j)$ is composed of the genes tagged by $T_j$. We use Gene Ontology (GO) (Ashburner *et al.*, 2000) as a reference library for annotating genes. For each gene, GO specifies the *molecular functions* associated with it, *biological processes* it takes part in, and *cellular components* it may be part of. The functional attributes in GO, known as GO terms, are organized hierarchically through *is a* and *part of* relationships. For example, 'regulation of stereoid biosynthetic process' is a 'regulation of stereoid metabolic process' and is part of 'stereoid biosynthetic process'. This hierarchy is abstracted using a directed acyclic graph (DAG). In this representation, if $T_i$ is a, or part of $T_j$, then $\mathcal{G}(T_i) \subseteq \mathcal{G}(T_j)$, *i.e.*, the genes associated with $T_i$ form a subset of genes associated with $T_j$. In this case, $T_j$ is said to be a *parent* of $T_i$. A term may have more than one parent, *i.e.*, $\mathcal{G}(T_i) \subseteq \mathcal{G}(T_j)$ and $\mathcal{G}(T_i) \subseteq \mathcal{G}(T_k)$ does not imply $\mathcal{G}(T_j) \cap \mathcal{G}(T_k) = \mathcal{G}(T_j) \cup \mathcal{G}(T_k)$. Furthermore, there is a unique $T_0 \in V_F$ with no parent, called *root*, such that $\mathcal{G}(T_0) = V_G$. In the rest of this section, we use a network of functional attributes with no constraints (*e.g.*, GO hierarchy) on function $\mathcal{G}$. We discuss the issue specifically relating to the GO hierarchy when addressing the implementation of NARADA.

We model networks of functional attributes using multigraphs. A multigraph is a generalized graph, where multiple edges are allowed between a single pair of nodes.

DEFINITION 2. **Functional Attribute Network.** *Given gene regulatory network $G(V_G, E_G)$, a set of functional attributes $V_F$, and functional annotation $\mathcal{A}(V_G, V_F) = \{\mathcal{F}, \mathcal{G}\}$, the corresponding functional attribute network $F(V_F, E_F)$ is a multigraph defined as follows. The set of functional attributes $V_F$ is also the set of nodes in $F$. Each node $T_i \in V_F$ contains a set of ports corresponding to the set of genes associated with $T_i$,* i.e., *$\mathcal{G}(T_i)$. Each multiedge $T_iT_j$ is a set of ordered port pairs (edges) $g_kg_\ell$, such that $g_k \in \mathcal{G}(T_i)$, $g_\ell \in \mathcal{G}(T_j)$, and $g_kg_\ell \in E_G$.*

The functional attribute network corresponding to the gene regulatory network in Figure 2(a) is shown in Figure 2(b). This multigraph model captures the context of each interaction accurately through the concept of ports. As illustrated in Figure 1, if a simple graph model is used, paths that do not exist in the gene network emerge in the functional attribute network. This is not possible in the multigraph model, since a *path* must leave a node from the port at which it enters the node.

DEFINITION 3. **Path.** *In functional attribute network $F(V_F, E_F)$, a path $\pi = \{(T_{i_1}, g_{j_1}), (T_{i_2}, g_{j_2}), ..., (T_{i_k}, g_{j_k})\}$ is an ordered set of node-port pairs such that (i) $T_{i_r} \neq T_{i_s}$ for $1 \leq r < s \leq k$ (nodes are not repeated), (ii) $g_{j_r} \in \mathcal{G}(T_{i_r})$ for $1 \leq r \leq k$, and (iii) $g_{j_r}g_{j_{r+1}} \in T_{i_r}T_{i_{r+1}} \in E_F$ for $1 \leq r < k$ (consecutive edges are connected through the same port). The length of $\pi$ is $|\pi| - 1 = k - 1$.*

In Figure 2(b), $\{(T_1, g_1), (T_2, g_3), (T_4, g_6)\}$ is a path but $\{(T_1, g_1), (T_2, g_4), (T_4, g_6)\}$ is not, since multiedge $T_1T_2$ does not contain the edge $g_1g_4$. Note that allowing $T_{i_1} = T_{i_k}$ and $g_{j_1} = gj_k$, we may also include cycles in this definition. According to the above definition, paths are characterized by ports. While analyzing regulatory pathways of functional attributes, however, we are interested in paths that are characterized by nodes in the functional attribute network. Clearly, such pathways may correspond to multiple paths in the functional attribute network. Therefore, we model them using *multipaths*.

DEFINITION 4. **Multipath.** *In functional attribute network $F(V_F, E_F)$, a multipath $\Pi = \{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$ is an ordered set of nodes such that (i) $T_{i_r} \neq T_{i_s}$ for $1 \leq r < s \leq k$, and (ii) there exist $g_{j_r} \in T_{i_r}$ for $1 \leq r \leq k$, such that $\{(T_{i_1}, g_{j_1}), (T_{i_2}, g_{j_2}), ..., (T_{i_k}, g_{j_k})\}$ is a path. The occurrence set $\mathcal{O}(\Pi)$ of $\Pi$ consists of all distinct paths that satisfy (ii) and each such path is called an occurrence of $\Pi$. The frequency of $\Pi$, $\phi(\Pi) = |\mathcal{O}(\Pi)|$, is equal to the number of occurrences of $\Pi$.*

We use the terms *pathway* and *multipath* interchangeably, to emphasize the biological meaning of a multipath. Allowing $T_{i_1} = T_{i_k}$, we also extend this definition to *multicycles*, occurrences of which correspond to cycles in the gene network. In Figure 2(b), $\{T_1, T_2, T_3\}$ (also denoted $T_1 \rightarrow T_2 \dashv T_3$ throughout this paper) is a multipath with frequency four. On the other hand, multipath $T_2 \dashv T_4 \rightarrow T_3$ does not exist in this network, *i.e.*, it has frequency zero, although multiedges $T_2 \dashv T_4$ and $T_4 \rightarrow T_3$ both exist. Note that the distinction between activator and inhibitor interactions is emphasized in this example for illustrative purposes, while it is omitted in the definition for simplicity. A multipath with *high* frequency is likely to be biologically interesting, since it corresponds to a regulatory pathway of functional attributes that recurs in various contexts in the underlying cellular organization. In order to quantify this biological significance, it is useful to evaluate frequency from a statistical perspective.

## Hardness of Significant Pathway Identification

Raw counts have long been used as a measure of significance – primarily because of the resulting algorithmic simplicity. This is a direct consequence of its *monotonicity* properties, namely that a subgraph (or substring/subset) of a frequent graph (or string/set) is itself frequent (Koyutürk *et al.*, 2006b). In identification of significantly overrepresented pathways of functional attributes, frequency alone does not provide a good measure of statistical significance. This is because, the degree distribution of gene regulatory networks and the distribution of the frequency of functional attributes are both highly skewed. Consequently, paths including functional attributes that are associated with high-degree genes (*e.g.*, molecular functions related to transcription) and those associated with many genes (*e.g.*, GO terms that are at coarser levels of GO hierarchy) are likely to dominate. For this reason, a statistical measure that takes into account these distributions is needed.

*Monotonicity of common statistical significance measures.* We identify the basic properties of a useful measure of statistical significance.

PROPOSITION 1. **Statistical Interpretability.** *Consider a set $\mathbf{X}$ of binary random variables and the set of corresponding observations $\mathbf{x}$, where $X = 1$ for $X \in \mathbf{X}$ corresponds to an observation*

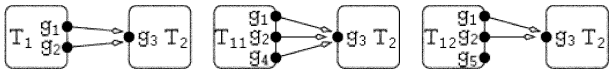


**Fig. 3.** Example illustrating that an interpretable measure of statistical significance is not monotonic with respect to GO hierarchy. GO terms $T_{11}$ and $T_{12}$ are parents of $T_1$. The regulatory effect of $T_{11}$ on $T_2$ is more significant than that of $T_1$, but the regulatory effect of $T_{12}$ is less significant.

*supporting a hypothesis. Let $f(\mathbf{X} = \mathbf{x})$ be a real-valued function, used to assess the statistical significance of the collection of observations defined by* $\mathbf{x}$*. Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be disjoint binary random variable sets,* i.e., $\mathbf{X} \cap \mathbf{Y} = \emptyset$*, and let* $\mathbf{x}$ *and* $\mathbf{y}$ *be the respective observation sets. A function $f$ is statistically interpretable if it satisfies the following conditions:*

*(i) If $y = 0 \; \forall \; y \in \mathbf{y}$, then $f(\mathbf{X} = \mathbf{x}) < f(\mathbf{X} \cup \mathbf{Y} = \mathbf{x} \cup \mathbf{y})$,*
*(ii) If $y = 1 \; \forall \; y \in \mathbf{y}$, then $f(\mathbf{X} = \mathbf{x}) > f(\mathbf{X} \cup \mathbf{Y} = \mathbf{x} \cup \mathbf{y})$.*

Here, without loss of generality, $f(\mathbf{X} = \mathbf{x}) < f(\mathbf{Y} = \mathbf{y})$ implies that $(\mathbf{X} = \mathbf{x})$ is a more interesting observation than $(\mathbf{Y} = \mathbf{y})$. More generally, the binary random variables characterize a pattern, and a larger set of these variables corresponds to a larger (or more general) pattern. This property simply states that additional positive (negative) observations should increase (decrease) our confidence that a pattern is interesting.

Most significance measures used in the analysis of discrete biological data are statistically interpretable. Consider, for example, the identification of significantly enriched GO terms in a set of genes. For a given term, the binary variables ($\mathbf{X}$), one for each gene ($X \in \mathbf{X}$), indicate whether the gene is associated with the term ($X = 1$). Adding a new gene ($\mathbf{Y} = \{Y\}$) to this set will improve the significance of enrichment ($f(\mathbf{x}) < f(\mathbf{x} \cup \mathbf{y})$) if the new gene is associated with the term ($Y = 1$). If not ($Y = 0$), the enrichment of the term in the new set will be less significant ($f(\mathbf{x}) > f(\mathbf{x} \cup \mathbf{y})$). Indeed, existing methods and statistical measures for this problem demonstrate this property (Hsiao *et al.*, 2005; Grossmann *et al.*, 2006).

Now we show that, in contrast to approximations that do not take into account the size of the sample space (*e.g.*, frequency), statistically interpretable measures of significance do not possess monotonicity.

THEOREM 1. *Let $f$ be a monotonically nondecreasing (nonincreasing) function,* i.e., *for any $\mathbf{X} \subseteq \mathbf{Z}$ and $\mathbf{x} \subseteq \mathbf{z}$, $f(\mathbf{X} = \mathbf{x}) \leq f(\mathbf{Z} = \mathbf{z})$ ($f(\mathbf{X} = \mathbf{x}) \geq f(\mathbf{Z} = \mathbf{z})$). Then $f$ is not statistically interpretable.*

PROOF. Without loss of generality, assume $f$ is nondecreasing. Let $\mathbf{Y}$ be a set of binary random variables, and $\mathbf{y}$ be a set of corresponding observations, such that $\forall y \in \mathbf{y}$, $y = 1$. Since $f$ is monotonically nondecreasing, we have $f(\mathbf{X} = \mathbf{x}) \leq f(\mathbf{X} \cup \mathbf{Y} = \mathbf{x} \cup \mathbf{y})$. This contradicts condition *(ii)* in Proposition 1. □

*Monotonicity with respect to GO hierarchy.* We now show that this result directly applies to the monotonicity of useful significance measures with respect to the GO hierarchy. Consider an ordered set of GO terms $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$. For any ordered set $\{g_{j_1}, g_{j_2}, ..., g_{j_k}\}$ such that $g_{j_r} \in \mathcal{G}(T_{i_r})$ for $1 \leq r \leq k$, define a binary random variable indicating the existence of the corresponding path in the underlying regulatory network. Clearly, the frequency of multipath $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$ is equal to the sum of the realizations of these random variables. Let $\mathbf{X}$ be the set of these random variables. Now, without loss of generality, consider pathway $\{T_P, T_{i_2}, ..., T_{i_k}\}$, such that $T_P$ is a parent of $T_{i_1}$, *i.e.*, $\mathcal{G}(T_{i_1}) \subset \mathcal{G}(T_P)$. Then, for each gene $g_P \in \mathcal{G}(T_P) \setminus \mathcal{G}(T_{i_1})$, there are multiple additional random variables, each for one of $\{g_P, g_{j_2}, ..., g_{j_k}\}$. Let $\mathbf{Y}$ be the set of these random variables. In this setting, the definition of statistical interpretability directly applies. If all paths of the sort $\{g_P, g_{j_2}, ..., g_{j_k}\}$ exist in the underlying regulatory network, then the pathway $\{T_P, T_{i_2}, ..., T_{i_k}\}$ is more significant than $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$. If none of them exist, then the pathway containing the child is more significant. Applying Theorem 1, we conclude that a statistically interpretable function, that quantifies the significance of the frequency of a multipath in the functional attribute network, cannot be monotonic with respect to GO hierarchy.

The example in Figure 3 illustrates this point. Here, both $T_{11}$ and $T_{12}$ are parents of $T_1$. Since all genes that are not in $T_1$ but in $T_{11}$ regulate $T_3$, the regulatory effect of $T_{11}$ on $T_3$ is more significant than that of $T_1$. Since none of the genes absent in $T_1$ but present in $T_{12}$ regulate $T_3$, the regulatory effect of $T_{12}$ on $T_3$ is less significant than that of $T_1$. Thus, any statistically interpretable measure $f$ should satisfy $f(T_{11} \rightarrow T_3) < f(T_1 \rightarrow T_3) < f(T_{12} \rightarrow T_3)$, which violates monotonicity. Note also that frequency, which is monotonically non-decreasing with respect to height (proximity to root) in GO hierarchy, is not statistically interpretable as $\phi(T_1 \rightarrow T_3) = \phi(T_{12} \rightarrow T_3)$.

This result can be interpreted as follows. GO hierarchy defines a combinatorial space of resolution for pathways of functional attributes. In other words, a pathway may be generalized or specialized by replacing a node (GO term) in the pathway with one of its ancestors or descendants in the GO DAG. Since this can be done for each node in the pathway, the size of this space is exponential in pathway length. However, as demonstrated above, the significance of a pathway fluctuates in this space. Consequently, all significant pathways cannot be efficiently identified using traditional inductive techniques, by starting from the highest (lowest) resolution in GO hierarchy and pruning out coarser (finer) terms in chunks.

Alternate approaches to this problem are necessary, not only in the context of significant pathway identification, but also other combinatorial problems in systems biology that involve hierarchical annotations. One possible approach is to develop a measure of statistical significance that admits a tight bound on the significance of a pathway in terms of the frequencies of pathways that are at a higher (lower) GO resolution. The discussion above clearly demonstrates that it is not straightforward to do so. Indeed, the statistical model we introduce in the next section does not easily lead to such tight bounds, since it emphasizes the *modularity* of a pathway to assess its significance. Consequently, in our implementation of NARADA, we use the most specific GO terms as the default resolution. Development of measures and methods that effectively prune out parts of the GO space remains an open problem.

*Monotonicity with respect to pathway length.* We apply Theorem 1 to the multipath space of a functional attribute network, *i.e.*, to the relationship between a multipath and its subpaths. As before, a multipath is represented by a set of binary random variables,

each corresponding to one of its potential occurrences. Without loss of generality, consider multipaths $\Pi_k = \{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$ and $\Pi_{k-1} = \{T_{i_1}, T_{i_2}, ..., T_{i_{k-1}}\}$. The random variables that represent $\Pi_k$ do not form a superset of those that represent $\Pi_{k-1}$. Rather, they are *extensions* of them, as defined below:

DEFINITION 5. **Extension.** *Given a set $\mathbf{X}$, an extension $\mathbf{Z}$ of $\mathbf{X}$, denoted $\mathbf{Z} \trianglerighteq \mathbf{X}$, is defined as follows. Each $X \in \mathbf{X}$, is attached to a subset $\mathbf{Z}_X \subseteq \mathbf{Z}$. Each $Z \in \mathbf{Z}$ is attached to exactly one $X \in \mathbf{X}$,* i.e., *for any $X_1, X_2 \in \mathbf{X}$, $\mathbf{Z}_{X_1} \cap \mathbf{Z}_{X_2} = \emptyset$.*

Each potential occurrence of $\Pi_k$ is a *superpath* of exactly one potential occurrence of $\Pi_{k-1}$ and there may be multiple such occurrences of $\Pi_k$ that correspond to a particular occurrence of $\Pi_{k-1}$. Therefore, the set of random variables that represent $\Pi_k$ form an extension of the set of random variables that represent $\Pi_{k-1}$.

PROPOSITION 2. **Statistical Interpretability w.r.t. Extension.** *Consider $\mathbf{X}$, $\mathbf{x}$, and $f(\mathbf{X} = \mathbf{x})$ as defined in Proposition 1. Let $Z \trianglerighteq X$ and let $\mathbf{z} \trianglerighteq \mathbf{x}$ be the respective observation set. A function $f$ is statistically interpretable with respect to extension if it satisfies the following conditions:*

*(i) If for all $x \in \mathbf{x}$ such that $x = 1$, $z = 0 \; \forall \; z \in \mathbf{z}_x$, then $f(\mathbf{X} = \mathbf{x}) < f(\mathbf{Z} = \mathbf{z})$,*

*(ii) If for all $x \in \mathbf{x}$ such that $x = 1$, $z = 1 \; \forall \; z \in \mathbf{z}_x$, then $f(\mathbf{X} = \mathbf{x}) > f(\mathbf{Z} = \mathbf{z})$.*

Each $x = 1$ corresponds to an occurrence of the corresponding pathway. Consequently, statistical interpretability with respect to extension of a pathway requires the following. If for all occurrences of $\Pi_{k-1}$, all corresponding potential occurrences of $\Pi_k$ exist in the network, then $\Pi_k$ is statistically more interesting than $\Pi_{k-1}$. If none of them occurs, then $\Pi_{k-1}$ is more interesting.

COROLLARY 1. *Let $f$ be a monotonically nondecreasing (nonincreasing) function with respect to extension,* i.e., *for any $\mathbf{Z} \trianglerighteq \mathbf{X}$ and $\mathbf{z} \trianglerighteq \mathbf{x}$, $f(\mathbf{X} = \mathbf{x}) \leq f(\mathbf{Z} = \mathbf{z})$ ($f(\mathbf{X} = \mathbf{x}) \geq f(\mathbf{Z} = \mathbf{z})$). Then $f$ is not statistically interpretable with respect to extension.*

The example shown in Figure 1 illustrates this result. In both of the scenarios shown in Figure 1(a) and (b), $\phi(T_1 \rightarrow T_2) = \phi(T_2 \rightarrow T_3) = 3$. In (a), $\phi(T_1 \rightarrow T_2 \rightarrow T_3) = 9$, *i.e.*, condition *(i)* in Definition 2 (all potential occurrences of $T_1 \rightarrow T_2 \rightarrow T_3$, given the occurrences of $T_1 \rightarrow T_2$, exist in the network), hence the pathway $T_1 \rightarrow T_2 \rightarrow T_3$ is more interesting than both $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_3$. In (b), on the other hand, $\phi(T_1 \rightarrow T_2 \rightarrow T_3) = 0$ (condition *(ii)* holds), so both $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_3$ are more interesting than $T_1 \rightarrow T_2 \rightarrow T_3$. This discussion motivates the statistical model we present in the next section.

## Statistical Model for Pathways of Functional Attributes

We present a novel statistical model for assessing the significance of the frequency of a multipath in a functional attribute network. In this approach, the "interestingness" of a pathway is associated with its *modularity*, *i.e.*, the significance of the coupling of its building blocks. In statistical terms, this is achieved by conditioning the distribution of the frequency (modeled as a random variable) of a pathway on the frequency of its subpaths (modeled as fixed parameters).

*Motivating example.* We illustrate the notion of the significance of coupling between regulatory interactions using the regulatory network and its corresponding functional attribute network shown in Figure 2. In this example, $\phi(T_1 \rightarrow T_2) = \phi(T_2 \dashv T_3) = \phi(T_2 \rightarrow T_4) = 2$, *i.e.*, regulatory interactions $T_1 \rightarrow T_2$, $T_2 \dashv T_3$, and $T_2 \rightarrow T_4$ occur twice. Furthermore, regulatory pathway (multipath in the functional attribute network) $T_1 \rightarrow T_2 \dashv T_3$ occurs four times, *i.e.*, $\phi(T_1 \rightarrow T_2 \dashv T_3) = 4$. Observe that, given the frequencies of $T_1 \rightarrow T_2$ and $T_2 \dashv T_3$, this is the maximum value $\phi(T_1 \rightarrow T_2 \dashv T_3)$ can take. In other words, any gene with annotation $T_2$, which is up-regulated by a $T_1$ gene, always down-regulates a $T_3$ gene. This observation suggests that, $T_1$ plays an indirect, but important role in the regulation of $T_3$. On the contrary, $\phi(T_1 \rightarrow T_2 \rightarrow T_4) = 2$, since gene $g_4$ with annotation $T_2$ up-regulates a $T_4$-gene ($g_6$), but it is not regulated by a $T_1$-gene. These observations suggest that the coupling between regulatory interactions $T_1 \rightarrow T_2$ and $T_2 \dashv T_3$ is stronger than the coupling between $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_4$. In other words, the pathway $T_1 \rightarrow T_2 \dashv T_3$ is more likely to be *modular*, compared to $T_1 \rightarrow T_2 \rightarrow T_4$.

We develop a statistical model that evaluates the modularity of regulatory pathways based on the coupling between their building blocks. For each pathway, our model assumes that the frequency of the building blocks of a pathway are known, *i.e.*, constitute the background distribution. We quantify the statistical significance of a pathway with the conditional probability of its frequency based on this background.

*Baseline model.* To quantify the significance of a pathway of shortest length (*i.e.*, a single regulatory interaction), we rely on a reference model that generates a functional attribute network. This model takes into account (i) the degree distribution of the underlying gene network, as well as (ii) the distribution of the number of genes associated with each functional attribute, based on the independent edge generation paradigm commonly used in modeling networks with arbitrary degree distribution (Chung *et al.*, 2003; Itzkovitz *et al.*, 2003). Note that this model is better suited to multigraphs than simple graphs (King, 2004). We refer to this model as the *baseline model*, and denote it $\mathcal{B}$.

The baseline model is defined by a set of parameters, and specifies the expected *multidegree* of each node in the functional attribute network. Here, the multidegree of a node in a multigraph refers to the number of multiedges incident to that node. Given gene regulatory network $G(V_G, V_E)$, functional attribute set $V_F$, and annotation $\mathcal{A}(V_G, V_F)$, the expected in-degree $\beta(T_i)$ and out-degree $\delta(T_i)$ of a functional attribute $T_i \in V_F$ are estimated as follows:

$$\hat{\beta}_i = \widehat{\beta(T_i)} = \sum_{T_j \in V_F} \phi(T_i T_j), \hat{\delta}_i = \widehat{\delta(T_i)} = \sum_{T_j \in V_F} \phi(T_j T_i), \tag{1}$$

where we denote the estimate of a parameter $x$ by $\hat{x}$. Note also that, if $f$ is a function of $x$, we use $f_i$ to denote $f(x_i)$ whenever appropriate. Given these parameters, $\mathcal{B}$ generates a functional attribute network as follows: there is a pool of *potential edges* that contains $\beta_i \delta_j$ potential edges between each pair of functional attributes $T_i$ and $T_j$. The size of the pool is given by: $m = \sum_{T_i, T_j \in V_F} \beta_i \delta_j$.A total of $n$ edges are drawn from this pool, independently and without replacement, where $n$ is equal to the number of edges in the observed functional attribute network, *i.e.*, $n = \sum_i \beta_i = \sum_j \delta_j$. Let $B_i = B(T_i)$ and $D_i = D(T_i)$ denote the random variables that

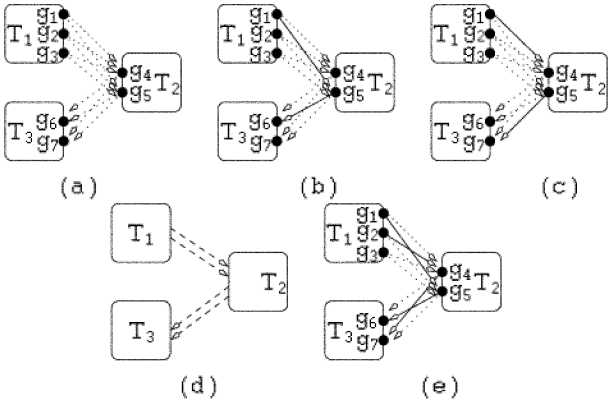


**Fig. 4.** Model testing whether the frequency of path $T_1 \rightarrow T_2 \rightarrow T_3$, given $\phi_{12}$ and $\phi_{23}$, is significant. (a) Pool of possible $T_1T_2$ and $T_1T_3$ edges. There are $\phi_1\phi_2 = 6$ and $\phi_2\phi_3 = 4$ possible $T_1T_2$ and $T_1T_3$ edges, respectively. (b) A possible pair of edges that corresponds to a path. (c) A possible pair of edges that does not correspond to a path. (d) $\phi_{12} = 2$ $T_1T_2$ edges and $\phi_{23} = 2$ $T_2T_3$ edges are randomly selected from the pool. (e) A possible configuration of selected edges. In this case, $\phi_{123} = 2$.

correspond to the in and out degrees of $T_i$ in the generated network. Then, we have

$$E[B_i] = \sum_j \beta_i\delta_j \frac{n}{m} = \beta_i \sum_j \delta_j \frac{\sum_\ell \beta_\ell}{\sum_{\ell,j} \beta_\ell\delta_j} = \beta_i \qquad (2)$$

and similarly $E[D_i] = \delta_i$. In other words, the expected values of multidegrees in the generated network mirror the specifications.

*Significance of a regulatory interaction.* Let $\Phi(\Pi)$ denote the random variable representing the frequency of pathway $\Pi$ in the generated functional attribute network. Clearly, $\Phi_{ij} = \Phi(T_iT_j)$ is a hypergeometric random variable with parameters $m$ (number of items), $\beta_i\delta_j$ (number of good items), $n$ (number of selected items), and $\phi_{ij}$ (number of selected good items) (Feller, 1968). Hence, the $p$-value of a regulatory interaction $T_iT_j$ in the observed network, *i.e.*, the probability of observing at least $\phi_{ij}$ interactions between genes associated with $T_i$ and genes associated with $T_j$, is given by

$$p_{ij} = P(\Phi_{ij} \geq \phi_{ij}|\mathcal{B}) = \sum_{\ell=\phi_{ij}}^{\min\{\beta_i\delta_j,n\}} \frac{\binom{\beta_i\delta_j}{\ell}\binom{m-\beta_i\delta_j}{n-\ell}}{\binom{m}{n}}. \qquad (3)$$

*Significance of a pathway.* We now present a statistical model to assess the statistical significance of a pathway of functional attributes, which assumes a background distribution based on the occurrence of the building blocks of a pathway. Let $\Pi_{i,k}$ denote the path $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$. For $1 < j < k$, we want to evaluate the significance of the coupling between pathways $\Pi_{1,j}$ and $\Pi_{j,k}$. In other words, we want to understand how strong a conclusion of the sort "If a gene $g_\ell \in \mathcal{G}(T_{i_j})$ is regulated through a chain of regulatory interactions characterized by $\Pi_{1,j}$, then this gene is likely to regulate a $T_{i_k}$ gene through pathway $\Pi_{j,k}$" (or vice versa) can be.

To achieve this, we assume a reference model, in which the frequency of pathways $\Pi_{1,j}$ and $\Pi_{j,k}$ is established *a-priori*. Let $\Phi_{i-k}$ and $\phi_{i-k}$ denote $\Phi(\Pi_{i,k})$ and $\phi(\Pi_{i,k})$, respectively. Then, the $p$-value of the coupling between $\Pi_{1,j}$ and $\Pi_{j,k}$ is defined as follows:

$$p_{1,j,k} = P(\Phi_{1,k} \geq \phi_{1,k}|\Phi_{1,j} = \phi_{1,j}, \Phi_{j,k} = \phi_{j,k}). \qquad (4)$$

Our model for the distribution of $\Phi_{1,k}$, given $\phi_{1,j}$ and $\phi_{j,k}$, is illustrated in Figure 4. Assume that a pool contains all possible occurrences of multipaths $\{T_{i_1}, T_{i_2}..., T_{i_j}\}$ and $\{T_{i_j}, T_{i_2}..., T_{i_k}\}$. Clearly, there are $m_{1,j} = \prod_{\ell=1}^{j} \phi_{i_\ell}$ and $m_{j,k} = \prod_{\ell=j}^{k} \phi_{i_\ell}$ potential occurrences of each multipath. This is shown in Figure 4(a). Now consider a pair of paths, one corresponding to a potential occurrence of $\Pi_{1,j}$, the other to $\Pi_{1,k}$. Such a pair corresponds to a path, *i.e.*, an occurrence of $\Pi_{1,k}$, only if the second path originates in the port in which the first one terminates. This is illustrated in Figure 4(b) and (c). Since there are $\phi_{1,j}$ and $\phi_{j,k}$ occurrences of $\Pi_{1,j}$ and $\Pi_{j,k}$, respectively, the problem is formulated as follows: we draw $\phi_{1,j}$ paths from $m_{1,j}$ potential occurrences of $\Pi_{1,j}$ and $\phi_{j,k}$ paths from $m_{j,k}$ potential occurrences of $\Pi_{j,k}$, forming $\phi_{1,j}\phi_{j,k}$ pairs. What is the probability that in at least $\phi_{1,k}$ of these pairs, the port on $T_j$ will be common?

We approximate this probability using our result on the behavior of dense subgraphs (Koyutürk *et al.*, 2006a) and Chvátal's bound on hypergeometric tail (Chvátal, 1979). In order to apply these results, we resolve dependencies assuming that the selected path pairs are independent from each other. Then, letting $q_j = 1/\phi_j$ be the probability that a given path pair will go through the same gene and $t_{1,j,k} = \phi_{1,k}/\phi_{1,j}\phi_{j,k}$ be the fraction of observed paths among all existing pairs, we obtain the following bound:

$$p_{1,j,k} \leq \exp(\phi_{1,j}\phi_{j,k}H_{q_j}(t_{1,j,k})), \qquad (5)$$

where $H_q(t) = t \log \frac{q}{t} + (1-t) \log \frac{1-q}{1-t}$ denotes weighted entropy. This estimate is Bonferroni-corrected for multiple testing, *i.e.*, it is adjusted by a factor of $\prod_{j=1}^{k} |\bigcup_{g_\ell \in T_{i_j}} \mathcal{F}(g_\ell)|$.

## NARADA: A Software for Identification of Significant Regulatory Pathways

Based on the above statistical model, we develop algorithms and a comprehensive software tool, NARADA, for projecting gene regulatory networks on the functional attribute domain.

The input to NARADA consists of three files: (i) a gene regulatory network, in which the source gene, target gene, and the mode of interaction are specified for each regulatory interaction, (ii) specification of the functional attributes and their relations (*e.g.*, Gene Ontology `obo` file), and (iii) annotation file that specifies the mapping between genes and functional attributes. NARADA currently handles three types of queries:

- $\mathtt{Q_1}$: Given a functional attribute $T$, find all significant pathways that are regulated by (originate from) genes that are associated with $T$.
- $\mathtt{Q_2}$: Given a functional attribute $T$, find all significant pathways that regulate (terminate at) genes that are associated with $T$.
- $\mathtt{Q_3}$: Given a sequence of functional attributes $T_{i_1}, T_{i_2}, ..., T_{i_k}$, find all occurrences of the corresponding pathway in the gene network and determine its significance.

A pathway is identified as being significant if its $p$-value is less than the critical $\alpha$-level, a user defined parameter.

NARADA delivers near interactive query response using a novel, biologically motivated pruning technique. We call a pathway *strongly significant* if all of its subpaths are significant. In biological

**Table 1.** Total number of significant pathways found by NARADA on *E. coli* transcription network for various path lengths.

| Pathway length | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| *All significant pathways* | 427 | 580 | 1401 | 942 |
| *Strongly significant pathways* | 427 | 208 | 183 | 142 |
| *Short-circuiting common terms* | 184 | 119 | 3 | 1 |

terms, a strongly significant pathway is likely to correspond to a significantly modular process, in which not only the building blocks of the pathway, but also its constituent building blocks are tightly coupled. In the context of queries implemented in NARADA, these subpaths are limited to those that originate from (terminate at) the query term. The option for searching strongly significant paths is also available in NARADA.

The main motivation in identification of significant regulatory pathways is understanding the crosstalk between different processes, functions, and cellular components. Therefore, functions and processes that are known to play a key role in gene regulation (*e.g.*, transcription regulator activity or DNA binding) may overload the identified pathways and overwhelm other interesting patterns. However, genes that are responsible for these functions are likely to bridge regulatory interactions between different processes (Lee *et al.*, 2002), so they cannot be ignored. For this reason, such GO terms are short-circuited, *i.e.*, if process $T_i$ regulates $T_j$, which is a key process in transcription, and $T_j$ regulates another process $T_k$, then the pathway $T_i \rightarrow T_j \rightarrow T_k$ is replaced with regulatory interaction $T_i \rightarrow T_k$.

## RESULTS AND DISCUSSION

We test NARADA comprehensively on the *E. coli* transcriptional network obtained from RegulonDB (Salgado *et al.*, 2006). The release 5.6. of this dataset contains 1364 genes with 3159 regulatory interactions. 193 of these interactions specify dual regulation. We separate these dual regulatory interactions as up and down regulatory interactions. We use Gene Ontology (Ashburner *et al.*, 2000) as a library of functional attributes. The annotation of *E. coli* genes is obtained from UniProt GOA Proteome (Camon *et al.*, 2004). Using the mapping provided by GO, the gene network is mapped to functional attribute networks of the three name spaces in GO. Mapping to the biological process space provides maximum coverage in number of genes annotated, 881 genes are mapped to one or more of 318 process terms. We discuss here results obtained by this mapping only. Results relating to molecular functions and cellular components, as well as comprehensive results on pathways of biological processes, are available at the NARADA website.

We use NARADA to identify all significant forward and reverse pathways of length 2 to 5. In order to identify these paths, we run queries $\mathtt{Q}_1$ and $\mathtt{Q}_2$ with a critical $\alpha$ of 0.01 on all 318 biological processes. The number of pathways obtained using combinations of the algorithmic options described in the previous section are shown in Table 1. On a Pentium M (1.6GHz) laptop with 1.21GB RAM the brute-force approach takes on average 0.5 seconds per query for path length 2, to 12 seconds per query for paths of length 5. For strongly significant paths, it takes less than 2 seconds per query for paths of length 5, while for shortcutting terms it is 8 seconds per query for paths of length 4. Strongly significant pathways, *i.e.*, those obtained by extending only significant pathways, compose a significant portion of the highly significant pathways. This observations suggests that significantly modular pathways are also likely to be composed of significantly modular building blocks.

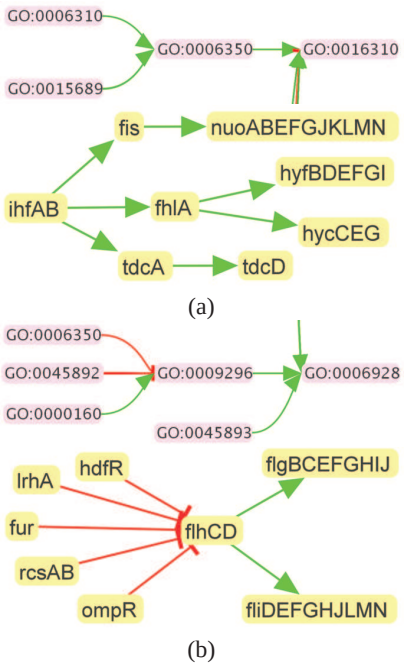


**Fig. 5.** Sample significantly overrepresented pathways in Ecoli transcription network. (a) DNA recombination → transcription → phosphorylation (b) transcription −| flagellum biogenesis → cell motility. The pathways in functional attribute space are shown on the upper panel, their occurrences in the gene network are shown on the lower panel.

*Discussion.* One of the prominent features of the detected significant pathways is that a large number of them begin with terms relating to transcriptional and translational regulation while ending in other cellular processes (Figure 5). This can be explained by the fact that the network consists of a set of transcription factor genes and set of genes regulated by them. Therefore, most of the regulatory pathways of length 3 or more have to begin at or flow through this set of genes annotated with processes relating to transcription, translation, and regulation thereof. Pathways involving other process terms occur with lower frequency, but most of them are highly significant.

Samples of pathways obtained are shown in Table 2. Some pathways like (sensory perception −| transcription → transport) occur frequently and may constitute a common mechanism for regulation of transport related activities. Parts of the significant pathways that regulate phosphorylation via genes involved in transcription and DNA recombination are shown in Figure 5(a). As genes involved in transcription are abundantly present in the network, part of the pathway (*DNA recombination → transcription*) occurs rarely (12 times) and is not significant, but in 6 of the 12 times it occurs, the genes involved in transcription regulate phosphorylation. The *fis* transcriptional regulator is responsible for regulation of *nuoA-N* operon (Wackwitz *et al.*, 1999), while the *fhlA* transcriptional

**Table 2.** Selection of significantly overrepresented pathways identified by NARADA on *E. coli* transcription network.

| Frequency | $p$-value | Pathway |
|---|---|---|
| 217 | 2.7E-49 | sensory perception $-\|$ transcription $\rightarrow$ transport |
| 64 | 7.1E-32 | regulation of translation $-\|$ DNA recombination $\rightarrow$ transport |
| 50 | 2.0E-24 | regulation of translation $-\|$ DNA recombination $-\|$ generation of precursor metabolites and energy |
| 45 | 1.1E-23 | molybdate ion transport $\rightarrow$ sensory perception $-\|$ metabolic process |
| 34 | 1.6E-8 | two-component signal transduction system (phosphorelay) $-\|$ transcription $\rightarrow$ sensory perception |
| 36 | 9.1E-8 | transcription $-\|$ flagellum biogenesis $\rightarrow$ chemotaxis |
| 37 | 6.7E-5 | two-component signal transduction system (phosphorelay) $\rightarrow$ transcription $\rightarrow$ cell motility |
| 6 | 6.2E-3 | sensory perception $-\|$ regulation of transcription, DNA-dependent $\rightarrow$ peptidoglycan catabolic process |
| 8 | 6.2E-3 | translation $-\|$ regulation of transcription, DNA-dependent $-\|$ detection of virus |
| 8 | 4.5E-3 | glcolysis $\rightarrow$ transcription $\rightarrow$ amino acid biosynthetic process |

activator regulates the *hyf* locus (Hopper *et al.*, 1994; Skibinski *et al.*, 2002). Indeed, it is observed that the integration host factor (*ihfA,ihfB*) affects the regulation of these phosphorylation related genes (*nuoA-N, hyf,hyc*) directly and indirectly (Hopper *et al.*, 1994; Nasser *et al.*, 2002).

In Figure 5(b), significant pathways that regulate cell motility are shown. This is part of a response to a query of type $\mathtt{Q}_2$. The *flhD* operon that encodes *flhC* and *flhD* has been shown to act as positive regulator of flagellar regulons(*fli, flg*) (Liu and Matsumura, 1994). The flagellar master operon *flhDC*, in turn, is tightly regulated at the transcriptional level by *rscAB, fur, ompR* (Ko and Park, 2000; Lehnen *et al.*, 2002; Francez-Charlot *et al.*, 2003). The output of NARADA captures this indirect regulation of flagellar expression perfectly.

*Case Study: Regulatory Network of Molybdate Ion Transport.* Figure 6 shows all significant paths of maximum length 3 regulated by molybdate ion transport. The genes associated with molybdate ion transport are *modE* and the operon *modABCD*, but it has been observed that the gene *modE* down-regulates the operon *modABCD* (McNicholas *et al.*, 1997), and the operon does not regulate any other gene. The three pathways at the bottom of the figure are the only significant paths of length 2 originating at molybdate ion transport. As can be seen on on the upper side of the figure (paths of length 3), molybdate ion transport promotes and suppresses various processes indirectly, through DNA-dependent regulation of transcription, two-component signal transduction system, and nitrate assimilation. It is important to note that direct regulation of these intermediate terms by molybdate ion transport is not significant by itself. By extending the search beyond pairwise interactions, NARADA is able to capture these significant indirect interactions successfully.

The paths of length 2 mirror the direct regulation of *moaABCDE* operon (McNicholas *et al.*, 1997) and *oppABCDF* operon (Tao *et al.*, 2005) by *modE*. Furthermore, *modE* indirectly regulates cytochrome complex assembly *ccm* operon (Overton *et al.*, 2006), electron transport *nap* operon (McNicholas and Gunsalus, 2002), nitrate assembly *nar* operon (Self *et al.*, 1999), and mitochondrial electron transport *nuo* operon (Bongaerts *et al.*, 1995; Overton *et al.*, 2006). All these indirect regulations occur through genes involved in respiratory nitrate reductase *narXL* (Tao *et al.*, 2005). In the RegulonDB network, we observe that *modE* indeed regulates *narL*, which regulates other genes. NARADA associates the mediation of *modE*'s

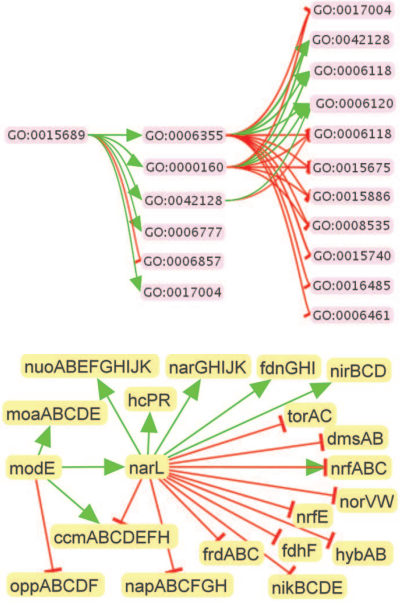


**Fig. 6.** Direct and indirect regulation of various processes by molybdate ion transport and the corresponding gene network.

regulatory effect on several other processes with the functional associations of *narL*.

An interesting observation is that, even though the regulation is mediated by the same gene, different biological processes associated with *narL* are found to mediate the regulation of different processes. Consider the paths molybdate ion transport $\rightarrow$ two component signal transduction $-|$ cytochrome complex assembly (1) and molybdate ion transport $\rightarrow$ nitrate assimilation $-|$ cytochrome complex assembly (2). Even though the underlying genes in both pathways are identical, the significance values assigned by NARADA to each of them is different (one is found to be significant while the other is not). Further inspection reveals that the regulatory interaction molybdate ion transport $\rightarrow$ two component signal transduction occurs only twice in the entire network, one of which, *modE* $\rightarrow$ *narL*, occurs in the context of (1). Similarly, two component signal transduction $-|$ cytochrome complex assembly occurs 9 times, 8 of which, *narL* $-|$ *nrfE,ccmABCDEFH*, occur in the context of

(1). On the other hand, molybdate ion transport → nitrate assimilation occurs 3 times in the complete network and is observed once in the context of (2), and only 8 of 15 occurrences of nitrate assimilation −| cytochrome complex assembly are associated with (2). Furthermore, there are 43 genes in the network that are associated with two component signal transduction, while there are 14 associated with nitrate assimilation. Consequently, statistical analysis suggests that a gene involved in two component signal transduction needs to be regulated by a molybdate ion transport to regulate cytochrome complex assembly. On the other hand, nitrate assimilation may regulate cytochrome complex assembly with and without the presence of molybdate ion transport gene regulating itself. Therefore, the modularity of the indirect suppression of cytochrome complex assembly by molybdate ion transport through two component signal transduction is found to be stronger than that through nitrate assimilation.

## CONCLUDING REMARKS

In this paper, we introduce the notion of statistically significant regulatory pathways of functional attributes and provide a formal framework for projecting regulatory networks from gene space to functional attribute space. We demonstrate the hardness of the resulting general problem in terms of non-monotonicity of interpretable statistical measures. We propose a statistical model for functional attribute networks that emphasizes the modularity of pathways by conditioning on its building blocks. We present a comprehensive software tool, NARADA, based on the proposed models and methods, and validate results obtained from the *E. coli* transcription network.

## REFERENCES


Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene Ontology: Tool for the unification of biology. the Gene Ontology consortium. *Nat Genet*, **25**(1), 25–29.

Bongaerts, J., Zoske, S., Weidner, U., and Unden, G. (1995). Transcriptional regulation of the proton translocating NADH dehydrogenase genes (nuoA-N) of Escherichia coli by electron acceptors, electron donors and gene regulators. *Mol Microbiol*, **16**(3), 521–534. Comparative Study.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen *et al.* (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, **32**(Database issue), 262–266.

Chung, F., Lu, L., and Vu, V. (2003). Spectra of random graphs with given expected degrees. *PNAS*, pages 6313–18.

Chvátal, V. (1979). The tail of the hypergeometric distribution. *Discrete Mathematics*, **25**, 285–287.

Feller, W. (1968). The hypergeometric series. In *An Introduction to Probability Theory and Its Applications*, volume 1, pages 41–45. Wiley, New York, 3rd edition.

Francez-Charlot, A., Laugel, B., Van Gemert, A., Dubarry, N., Wiorowski, F., Castanie-Cornet *et al.* (2003). RcsCDB His-Asp phosphorelay system negatively regulates the flhDC operon in Escherichia coli. *Mol Microbiol*, **49**(3), 823–832.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J Comput Biol*, **7**(3-4), 601–620.

Gamalielsson, J., Nilsson, P., and Olsson, B. (2006). A GO-based method for assessing the biological plausibility of regulatory hypotheses. In *International Conference on Computational Science (2)*, pages 879–886.

Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2006). An improved statistic for detecting over-represented gene ontology annotations in gene sets. *RECOMB'06*, pages 85–98.

Hopper, S., Babst, M., Schlensog, V., Fischer, H. M., Hennecke, H., and Bock, A. (1994). Regulated expression in vitro of genes coding for formate hydrogenlyase components of Escherichia coli. *J Biol Chem*, **269**(30), 19597–19604.

Hsiao, A., Ideker, T., Olefsky, J. M., and Subramaniam, S. (2005). VAMPIRE microarray suite: A web-based platform for the interpretation of gene expression data. *Nucleic Acids Research*, **33**(Web Server issue).

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, **19**(17), 2271–2282.

Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., and Alon, U. (2003). Subgraphs in random networks. *Physical Review E*, **68**, 026127.

King, O. D. (2004). Comment on "Subgraphs in random networks". *Physical Review E*, **70**, 058101.

Ko, M. and Park, C. (2000). H-NS-Dependent regulation of flagellar synthesis is mediated by a LysR family protein. *J Bacteriol*, **182**(16), 4670–4672.

Koyutürk, M., Grama, A., and Szpankowski, W. (2006a). Assessing significance of connectivity and conservation in protein interaction networks. In *RECOMB'06*, pages 45–59.

Koyutürk, M., Kim, Y., Subramaniam, S., Szpankowski, W., and Grama, A. (2006b). Detecting conserved interaction patterns in biological networks. *Journal of Computational Biology*, **13**(7), 1299–1322.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber *et al.* 2002). Transcriptional regulatory networks in S. cerevisiae. *Science*, **298**(5594), 799–804.

Lehnen, D., Blumer, C., Polen, T., Wackwitz, B., Wendisch, V. F., and Unden, G. (2002). LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in E. coli. *Mol Microbiol*, **45**(2), 521–532. Comparative Study.

Liu, X. and Matsumura, P. (1994). The FlhD/FlhC complex, a transcriptional activator of the E. coli flagellar class II operons. *J Bacteriol*, **176**(23), 7345–7351.

McNicholas, P. M. and Gunsalus, R. P. (2002). The molybdate-responsive E. coli ModE transcriptional regulator coordinates periplasmic nitrate reductase (nap) operon expression with nitrate and molybdate availability. *J Bacteriol*, **184**(12), 3253–3259.

McNicholas, P. M., Rech, S. A., and Gunsalus, R. P. (1997). Characterization of the ModE DNA-binding sites in the control regions of modABCD and moaABCDE of Escherichia coli. *Mol Microbiol*, **23**(3), 515–524.

Nasser, W., Rochman, M., and Muskhelishvili, G. (2002). Transcriptional regulation of fis operon involves a module of multiple coupled promoters. *EMBO J*, **21**(4), 715–724.

Overton, T. W., Griffiths, L., Patel, M. D., Hobman, J. L., Penn, C. W., Cole, J. A., and Constantinidou, C. (2006). Microarray analysis of gene regulation by oxygen, nitrate, nitrite, FNR, NarL and NarP during anaerobic growth of Escherichia coli: new insights into microbial physiology. *Biochem Soc Trans*, **34**(Pt 1), 104–7.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sánchez-Solano, F., Santos-Zavaleta, *et al.* (2006). RegulonDB (version 5.0): E. coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *NAR*, **34**.

Self, W. T., Grunden, A. M., Hasona, A., and Shanmugam, K. T. (1999). Transcriptional regulation of molybdoenzyme synthesis in E. coli in response to molybdenum: ModE-molybdate, a repressor of the modABCD (molybdate transport) operon is a secondary transcriptional activator for the hyc and nar operons. *Microbiology*, **145 ( Pt 1)**(NIL), 41–55.

Skibinski, D. A. G., Golby, P., Chang, Y.-S., Sargent, F., Hoffman, R., Harper, R.*et al.* (2002). Regulation of the hydrogenase-4 operon of E. coli by the sigma(54)-dependent transcriptional activators FhlA and HyfR. *J Bacteriol*, **184**(23), 6642–53.

Tao, H., Hasona, A., Do, P. M., Ingram, L. O., and Shanmugam, K. T. (2005). Global gene expression analysis revealed an unsuspected deo operon under the control of molybdate sensor, ModE protein, in E. coli. *Arch Microbiol*, **184**(4), 225–33.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, **36**(5), 492–496. Comparative Study.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X. *et al.* (2004). Global mapping of the yeast genetic interaction network. *Science*, **303**(5659), 808–813.

Wackwitz, B., Bongaerts, J., Goodman, S. D., and Unden, G. (1999). Growth phase-dependent regulation of nuoA-N expression in E. coli K-12 by the Fis protein: upstream binding sites and bioenergetic significance. *Mol Gen Genet*, **262**(4-5), 876–883.