ANNOTATING PATHWAYS IN INTERACTION NETWORKS

JAYESH PANDEY*, MEHMET KOYUTÜRK†, WOJCIECH SZPANKOWSKI AND ANANTH GRAMA

Department of Computer Science, Purdue University,
West Lafayette, IN 47907, USA
**E-mail: jpandey@cs.purdue.edu*

Integrating molecular interaction data with existing knowledge of molecular function reveals mechanisms that underly cellular organization. We present NARADA, a software tool that implements a comprehensive analysis suite for functional annotation of pathways. NARADA takes as input a species-specific molecular interaction network and annotation of biomolecules in the network and provides the user with a set of pathways composed of functional attributes, which may be thought of pathway templates in the functional annotation space that recur in various contexts (different groups of specific molecules with similar functional annotation patterns) in the molecular interaction network. NARADA has its underpinnings in formal statistical measures of significance, and algorithmic bases for performance. Comprehensive evaluation on the *E. coli* transcriptional regulation and protein-protein interaction data demonstrate NARADA's ability to detect known, as well as novel pathways.

1. Introduction

Network models are commonly used to abstract biomolecular interactions. Recent research has focused on identifying common patterns in these networks, within and across species, with the expectation that such patterns reveal evolutionary design principles that underly cellular organization. Indeed, coherent topological motifs (*e.g.*, feedback and feed-forward loops) and their constituent molecules are observed to recur significantly in the protein-protein interaction and transcriptional regulatory networks of model organisms [1]. Comparative analysis of extant networks also suggests that modular subcomponents of these networks are likely to be conserved together [2,3].

These observations support the hypothesis that the organizational principles that underly interaction networks may be represented in the form of functional (sub)networks – "rules" or "templates" that recur in various contexts in the functional organization of the cell. The underlying problem of generalizing from molecular annotations, provided by libraries such as Gene Ontology [4], to subnetwork annotations is important – and forms the technical challenge addressed in this paper. Preliminary studies show that such annotations can indeed be derived;

†Present address: *Department of Electrical Engineering & Computer Science, Case Western Reserve University, Cleveland, OH, USA*

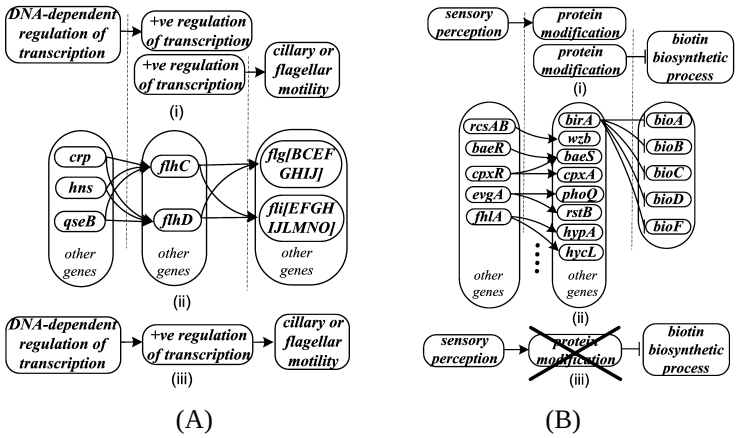


Fig. 1. (A) From interactions between functional attributes to pathways of functional attributes. (B) Significant pairwise interactions between functional attributes do not necessarily imply indirect paths.

however, they do not provide an automated methodology, or a comprehensive analytical (statistical) basis for the annotations [5–8]. Schwikowski *et. al.* [5] predict functions of proteins in *S. cerevisiae* protein-protein interaction network by hypothesizing that proteins of known function and cellular location tend to cluster together. Lee *et. al.* [6] study the *S. cerevisiae* transcriptional regulatory network with a view to understanding relationships between functional categories of genes. They observe 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 bind to several genes that regulate metabolism, environmental response, and development. Tong *et. al.* [7] 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.

These results are limited to case-specific studies that generally focus on validation or evaluation of results through simple statistical analyses – yet they provide significant insights. Generalizing these observations allows identification of standardized pathways, creation of reference databases of direct and indirect interactions between various processes, and projecting existing knowledge of model organisms to other species. What is lacking is a comprehensive set of tools that combine these sources of data (molecular annotations and interactions) to identify significantly overrepresented patterns of interaction through reliable statistical modeling with a formal computational basis.

In recent work [9], we explore the statistical and algorithmic underpinnings of this problem. In this paper, we describe a comprehensive toolkit, NARADA,

for pathway annotation. NARADA can be applied to diverse abstractions (*e.g.*, gene regulatory networks, protein-protein interaction networks), and can use as reference node annotations any user-specified ontology. Users can specify functional categories of interest, query for statistically over-represented pathways in terms of these functional categories, visually manipulate and inspect these pathways, and view reflections of these pathways in "molecular" (*i.e.*, gene network) and "functional" (*i.e.*, network of functional attributes) space. NARADA evaluates the statistical significance of pathways based on a novel statistical model, which emphasizes the modularity of pathways by conditioning on the frequency of their building blocks. NARADA is implemented in Java and is available as a web applet, as well as a standalone application at http://www.cs.purdue.edu/~jpandey/narada.

2. Models

Molecular interactions are abstracted using various network models. Regulation of gene expression, for example, is commonly modeled using Boolean networks [10]. Protein-protein interactions (PPIs), on the other hand, represent various forms of physical association between proteins, including modification, transport, and complex formation [11]. NARADA is designed to handle different types of networks and different sources of data in a unifying framework. In this section, for the sake of clarity, we present the mathematical underpinnings of NARADA in the context of gene regulatory networks, and focus on identification of regulatory pathways.

The basic approach to integrating existing knowledge of gene networks and functional annotation is to (i) project nodes from the *gene space* onto the *functional attribute space*, and (ii) find significant pathways in the functional attribute space. The first step is accomplished using a reference node annotation library. A simple method for accomplishing the second step is to identify statistically abundant (significant) pairs of interacting functional attributes. For example, in Figure 1(A), the *E. coli* transcription network contains 36 activator interactions between 2 genes that take part in *positive regulation of transcription* and 18 genes that are involved in *cillary or flagellar motility*. This observation may be abstracted as a *rule* that characterizes the regulatory relationship between these two processes: *positive regulation of transcription* up-regulates *cillary or flagellar motility* in *E. coli*. Indeed, this approach is used to understand the functional organization of *S. cerevisiae* synthetic genetic array [7] and transcriptional regulatory networks [6].

Statistically significant pathway annotations cannot be directly composed from constituent pairwise annotations. This is because, each interaction between a pair of functional attributes is within a specific context (a different pair of genes) in

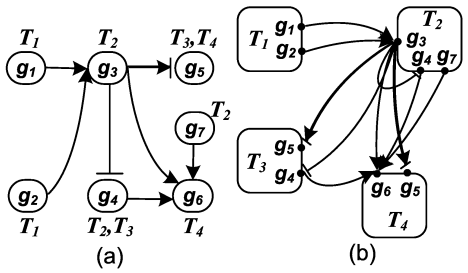


Fig. 2. (a) A sample gene regulatory network and the functional annotation of the genes in this network. Each node represents a unique gene and is tagged by the set of functional attributes attached to that gene. Activator interactions are shown by regular arrows, repressor interactions are shown by 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.

the network. This is illustrated in Figure 1. In both (A) and (B), the two regulatory interactions shown on the panel (i) are significantly frequent in the gene network. In Figure 1(A), the genes involved in *positive regulation of transcription* shown in panel (ii) are common to both interactions and the combined pathway shown in the panel (iii) is frequent. On the other hand, in Figure 1(B), the set of genes involved in *protein modification* (in panel (ii)) are different for the two interactions, so the combined pathway (in panel (iii)) does not exist in the gene network at all! However, a method that relies on assessment of only pairwise interactions would identify indirect regulation of *biotin biosynthetic process* by *sensory perception* through *protein modification* as a significant pathway, which is not a conclusion that is supported by available data.

Data Model. A gene network is modeled as a labeled directed graph with nodes representing genes and edges representing regulatory interactions. Each edge is associated with a type that specifies the mode of regulation (activation, repression or dual). Each gene in the network is associated with a set of *functional attributes*, which provide functional annotations of the gene. Without loss of generality, we use Gene Ontology (GO) [4] to annotate the genes in the network.

Given a gene network and annotation, the corresponding *functional attribute network* is defined as follows: each functional attribute T_i is represented by a *multinode*, which contains a set of *ports*, each corresponding to a gene g_j that is associated with T_i. The frequency $\phi(T_i)$ of a functional attribute is equal to the number of genes that are associated with T_i. Each *multiedge* T_iT_j corresponds to a set of edges g_kg_l in the gene network, such that g_k is associated with T_i and g_l is associated with T_j. The frequency $\phi(T_iT_j)$ of a multiedge is equal to the number of such edges in the gene network. A *multipath* of length k is a sequence of k distinct functional attributes (multinodes), $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$. A sequence of

genes $\{g_{j_1}, g_{j_2}, ..., g_{j_k}\}$ is an occurrence of multipath $\{T_{i_1}, T_{i_2}, ..., T_{i_k}\}$ if each g_{j_ℓ} is associated with the corresponding T_{i_ℓ} and there is an edge from g_{j_ℓ} to $g_{j_{\ell+1}}$ in the gene network for each ℓ. The frequency $\phi(\{T_{i_1}, T_{i_2}, ..., T_{i_k}\})$ of a multipath is the number of occurrences of that multipath in the gene network.

A sample gene network and its corresponding functional attribute network is shown in Figure 2. In Figure 2, the frequency of multipath $T_1 \rightarrow T_2 \dashv T_3$ is 4.

Statistical Model. 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). Note that, in this approach, statistical significance is used as an indicator of the modularity of a pathway in the functional annotation space, *i.e.*, the hypothesis that is tested here is that a pathway of functional attributes corresponds to a design template that is conserved and rediscovered through evolution [12]. Thefore, the statistical significance of a pathway should be interpreted as the likelihood that the observed pattern is biologically relevant (in Kitano's [12] terms, it may have a place in the "periodic table" of functional regulatory circuits), rather than being a measure of the pattern's biological relevance or importance.

A single interaction is the shortest pathway in a functional attribute network. We evaluate the significance of single interaction by taking into account the frequency of each functional attribute and the degree distribution of the gene network. For each functional attribute T_i, its expected in-degree β_i and out-degree δ_i are specified. Then, edges are generated by randomly selecting $n = \sum_i \beta_i = \sum_j \delta_j$ edges from $m = \sum_{T_i, T_j \in V_F} \beta_i \delta_j$ potential edges, where each of the $\beta_i \delta_j$ potential edges are between T_i and T_j. Letting $\Phi_{ij} = \Phi(T_i T_j)$ be frequency of $T_i T_j$ in the random model, we observe that Φ_{ij} is a hypergeometric random variable and obtain $p_{ij} = P(\Phi_{ij} \geq \phi_{ij}) = \sum_{\ell=\phi_{ij}}^{\min\{\beta_i\delta_j, n\}} \binom{\beta_i\delta_j}{\ell}\binom{m-\beta_i\delta_j}{n-\ell} / \binom{m}{n}$.

Now 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 $\Pi_{j,k}$. Our reference model assumes that 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 $\phi(\Pi_{i,k})$, the random variable that represents the frequency of pathway $\Pi_{i,k}$ and its observed value, respectively. Then, the p-value of the coupling between $\Pi_{1,j}$ and $\Pi_{j,k}$ is defined as $p_{1,j,k} = P(\Phi_{1,k} \geq \phi_{1,k} | \Phi_{1,j} = \phi_{1,j}, \Phi_{j,k} = \phi_{j,k})$.We approximate this value using Chvátal's bound on hypergeometric tail [13] to obtain $p_{1,j,k} \leq \exp(\phi_{1,j}\phi_{j,k} H_{q_j}(t_{1,j,k}))$,where $t_{1,j,k} = \phi_{1,k}/\phi_{1,j}\phi_{j,k}$, $q_j = 1/\phi_j$ (ϕ_j is frequency of term T_j), and $H_q(t) = t \log \frac{q}{t} + (1-t) \log \frac{1-q}{1-t}$ denotes divergence. This estimate is Bonferroni-corrected for multiple testing, *i.e.*, it is adjusted by a factor of $\prod_{j=1}^{k} |\bigcup_{g_\ell \in T_{i_j}} \mathcal{F}(g_\ell)|$.

3. Methods and Features

NARADA is implemented in Java, and can be run as a web applet or a standalone application. It requires an installation of Java Runtime Environment version 1.4.2_14 (update 14) or later and has been tested to run of windows and linux platform. The base application ERIS is based on Cytoscape [14]. This software framework allows for development of sophisticated visualization and analysis functions through java-based plugins. The user manual and source code for NARADA are available at http://www.cs.purdue.edu/homes/jpandey/narada.

Query Interface. Currently, NARADA supports three classes of queries:

- $\mathtt{Q}_1$: Given a functional attribute T, find all significant pathways that are regulated by (originate at) 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.

A pathway is identified as being significant if its p-value is less than a user-specified level, α. Pathways do not have repeated internal nodes, but cycles (feedback loops) are allowed, *i.e.*, the output to queries $\mathtt{Q}_1$ ($\mathtt{Q}_2$) may include a pathway that terminates (originates) at the queried term itself, provided each occurrence of the cycle corresponds to a cycle in the gene network.

Algorithms. For a given term, we perform an enumerative search on the functional attribute network (without explicitly constructing the network) starting from the node that corresponds to the query term. For queries of type $\mathtt{Q}_1$ ($\mathtt{Q}_2$), the search proceeds forwards (backwards) with respect to edge direction. Consequently, the output is a tree that is rooted at the query term.

During the course of the search process, the significance of each pathway is tested as follows: if the length of the pathway is one, *i.e.*, it is a multiedge, its significance is evaluated with respect to the baseline model. Otherwise, assume that we are trying to extend pathway $\{T_{i_1}, ..., T_{i_{k-1}}\}$ by adding multiedge $T_{i_{k-1}}T_{i_k}$, where T_{i_1} is the query term. We condition the significance of pathway $\{T_{i_1}, ..., T_{i_k}\}$ on the frequency of pathways $\{T_{i_1}, ..., T_{i_{k-1}}\}$ and $T_{i_{k-1}}T_{i_k}$. The motivation behind this is as follows: if the regulatory effect of $T_{i_{k-1}}$ on T_{i_k} is significantly coupled with pathway $\{T_{i_1}, ..., T_{i_{k-1}}\}$, *i.e.*, a significant number of its occurrences in the network are likely to be preceded by this pathway, then this may correspond to a *rule* that characterizes the regulation of T_{i_k} through a chain of regulatory interactions specified by pathway $\{T_{i_1}, ..., T_{i_k}\}$.

For queries from class $\mathtt{Q}_3$, consider the sequence of functional attributes

$T_{i_1}, T_{i_2}, ..., T_{i_k}$. For such a query, NARADA finds all occurrences of the pathways $\{T_{i_1}, ..., T_{i_j}\}$, $\{T_{i_j}, ..., T_{i_k}\}$ and $\{T_{i_1}, ..., T_{i_k}\}$. To find all occurrences of a pathway, for each functional attribute, the genes that bridge the previous and next node in the sequence are identified. Then, the frequencies of these pathways are used to compute the significance as described in the previous section. By mapping all genes that are identified to the gene network, NARADA displays all occurrences of the pathway in the gene network.

Performance Enhancement and Heuristics. A major limitation of the algorithm above is that it is brute force and its time complexity is exponential in k (length of the path). The longest pathway that is of interest is a user-defined parameter in NARADA. Since pathways of biological significance are expected to be fairly short, the practical constraints posed by this exponential complexity are somewhat mitigated. However, since there exist several genes that are attached to many functional attributes and vice versa, the branching factor of the search process is quite large. For this reason, pruning heuristics that render significant pathway identification tractable for very large networks and longer pathways are still necessary. In NARADA, various heuristics that exploit *a priori* biological knowledge are implemented to accelerate the search process. We outline these heuristics below. We also note that development of efficient heuristics that integrate syntactic and semantic information remains an important open problem.

Gene Ontology hierarchy: The current release of NARADA uses Gene Ontology (GO) [4] as the default reference library for annotations. NARADA's default behavior in handling this hierarchy is to use the most specific GO term on each branch of GO hierarchy for each gene. In other words, if terms T_i and T_j are attached to gene g_ℓ and if T_j is a parent of T_i in GO hierarchy (*i.e.*, either T_i *is a* T_j or T_i is *part of* T_j), then only T_i is considered in the functional attribute network. The user is allowed to alter this behavior by selecting to annotate the genes using any specific level of the hierarchy. Each query can also be refined by moving a term in the query up or down the GO hierarchy.

Strongly significant pathways: NARADA delivers near interactive query response using a biologically motivated pruning technique. We call a pathway *strongly significant* if all of its subpaths are significant. In biological terms, a strongly significant pathway is likely to correspond to a significantly modular process, in which not only the building blocks of the pathway are significant, but are also tightly coupled. This makes it possible to extend pathway length without significant re-computation. For queries of type $\mathtt{Q}_1$ and $\mathtt{Q}_2$, the option for searching strongly significant paths is available in NARADA.

Short-circuiting common terms: 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 [6], 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 $T_i \rightarrow T_k$.

Interface. A user interface with comprehensive functionality and visualization capabilities is available in NARADA. The visualization infrastructure is built using an open source library, Prefuse [15], which provides standard graph visualization functions. The current version of NARADA can handle large networks with thousands of genes and annotations. The graph views (for both gene and functional attribute space) support pan, drag, zoom, and standard layout functionalities, search by node name and node link-outs to biological databases. Screenshots from NARADA are shown in Figure 3.

The input to NARADA consists of three files:

- A molecular interaction network, in which interacting molecules and type of interaction are specified using the simple interaction file `(sif)` format [14]. Multiple networks can be loaded simultaneously, NARADA creates separate visualizations for each. These networks may belong to different organisms.
- Specification of the functional attributes and their relations (*e.g.*, Gene Ontology (GO) `obo` file). Currently, only one attribute set can be used in one session.
- Annotation file that specifies the mapping between nodes and their functional attributes. Multiple annotation files can be loaded to provide mapping for one or more networks.

Detecting pathway annotations: The interface to query significant pathways originating (or terminating) at a functional attribute allows the user to specify α parameter, the limit on pathway length, and a flag indicating whether the search is limited to strongly significant pathways. The result of a query $\mathtt{Q}_1(\mathtt{Q}_2)$ is displayed as a tree. Each path from the root to a leaf represents a significantly overrepresented pathway. The p-value of each pathway is stored at the corresponding leaf. Each pathway can also be separately viewed in a GO Path frame, which offers the user the ability to move up/down the GO hierarchy for any node in the pathway. Moreover, this interface also allows the user to view all occurrences of the pathway in the gene network. It is also possible to submit a single query to NARADA to run queries of type $\mathtt{Q}_1, \mathtt{Q}_2$ for all functional attributes in bulk. The results of such a query can be directly written to an output file. To query all occurrences of

a specified pathway of attributes (query type $\mathtt{Q}_3$), the user enters the sequence of GO terms, specifies the edges along with their types (*e.g.*, mode of regulation), and the output can then be explored through the GO path view.

4. Results and Discussion

We run NARADA on the *E. coli* transcriptional network and *E.coli* protein interaction network to identify core functional pathways that underly cellular regulation and signaling in *E.coli*.

We obtain the *E. coli* transcriptional network (TrN) from RegulonDB [16]. The release 5.6 of this dataset contains 1363 genes with 3159 regulatory interactions. The *E. coli* protein interaction network (PIN) is obtained from DIP [17]. The latest release (20070219) of this dataset contains 1841 proteins with 6958 interactions. We use Gene Ontology [4] as a library of functional attributes. The annotation of *E. coli* genes and proteins is obtained from the UniProt GOA Proteome 48.0 release [18]. Using the default 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 or proteins annotated. In the TrN 904 genes are mapped to one or more of 340 process terms, while for the PIN 793 proteins are mapped to one or more of 343 process terms. We discuss here results obtained by this mapping only.NARADA is equally useful for the molecular function branch of the GO, with results like *transcription factor binding* $->$ *ATP binding* $->$ *electron carrier activity*. Results relating to molecular functions and cellular components, as well as comprehensive results on pathways of biological processes for both networks, are available at http://www.cs.purdue.edu/homes/jpandey/narada/.

We use NARADA to identify all significant pathways of length 2 to 5. In order to identify these paths, we run queries $\mathtt{Q}_1$ (and $\mathtt{Q}_2$ for transcription network) with a critical α of 0.01 on all annotated biological processes. The number of pathways obtained using combinations of the algorithmic options described in the previous section are shown in Table 1. These results differ from previous results in [9] on account of better annotation which affects the bonferroni-correction. On a Pentium M (1.6GHz) laptop with 1.21GB RAM identification of all significantly over-represented pathways takes average of 1.2s per query for TrN, and 8s per query for PIN, for upto path length 5 and 4 respectively. For strongly significant paths, it takes less than 0.5s per query in the TrN, and less than 2s per query in the PIN for paths of length upto 5. Strongly significant pathways, *i.e.*, those obtained by extending only significant pathways, compose an important part of the highly significant pathways. This observation suggests that significantly modular pathways are also likely to be composed of significantly modular building blocks. In

Table 1. Total number of significant pathways identified by NARADA for various path lengths.

E. coli network	algorithm	2	3	4	5
transcriptional	*All significant pathways*	213	1404	3472	2251
	Strongly significant pathways	213	210	248	148
	Short-circuiting common terms	445	422	371	38
protein interaction	*All significant pathways*	208	3533	53486	-
	Strongly significant pathways	208	699	4196	36266

the TrN after short-circuiting terms related to transcription, translation, and regulation thereof, identification of all significant paths takes 0.9s per query for paths of length 5. Note that a short-circuited path of length 5 might actually correspond to a path of length upto 9 with hidden (short-circuited) nodes.

Sample results: Parts of the significant pathways that regulate phosphorylation via genes involved in transcription and DNA recombination are shown in Figure 4(a). As genes involved in transcription are abundantly present in the network, part of the pathway (*DNA recombination* $\rightarrow$ *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 and the complete pathway occurs 38 times ($p < 4 \times 10^{-15}$). The *fis* transcriptional regulator is responsible for regulation of *nuoA-N* operon [19], while the *fhlA* transcriptional activator regulates the *hyf* locus [20]. Indeed, it is observed that integration host factor (*ihfA,ihfB*) affects the regulation of these phosphorylation related genes (*nuoA-N, hyf,hyc*) directly and indirectly [20].

Figure 4 (b) shows a significant pathway that is composed of *translation*, *DNA replication* and *protein folding*, as well as the corresponding proteins and their interactions in the PPI network. This pathway recurs 20 times ($p < 3.6 \times 10^{-3}$) in the PPI network. Proteins involved in *DNA replication* are abundantly present in the network, but are connected to proteins involved in *protein folding* only 8 times. 5 out of these 8 interactions are preceded by proteins involved in translation. The dnaK chaperone system, consisting of dnaK, dnaJ and grpE are involved in remodeling and refolding of proteins, with cbpA functioning as a dnaJ-like co-chaperon [21,22]. dnaA involved in DNA replication activity is protected by dnaK from reaching a self-aggregate inactive form [23]. Most of the other proteins involved in translation form part of ribosomal assembly or translational elongation factor activity [24].

A global view of E.coli functional regulatory network: A summary of all significant pathways identified on *E. coli* transcription network is shown in Figure 5. This view provides a mapping of the *E.coli* transcriptional network to the biological process space of Gene Ontology. In the figure, the top 20% significant path-

ways for pathway lengths 2 to 4 are shown. The edges that constitute significant pathways of length 2, 3, and 4 are shown using solid, dashed, and dotted lines, respectively. This results in a connected network of 71 functional attributes. In the figure, the font size of each GO term is proportional to its degree in this network, and thickness of an edge is proportional to its significance (or significance of the pathway it is a part of). As seen in the figure, this network is clustered into various fundamental processes. A large subnetwork consists of processes related to response to stress and stimulus, DNA repair, and negative regulation of transcription. This subnetwork is mostly composed of down-regulatory interactions. A second important group of processes that are tightly coupled in this network relates to cell motility, cytochrome assembly, flagellum and positive regulation of transcription. These processes are mostly connected via up-regulatory interactions. Observe that the regulatory interactions in these *local* subnetworks correspond to significant pathways of length 2, *i.e.*, they are direct regulatory interactions, but they also may be parts of significant indirect pathways.

The edges that are part of significant indirect pathways (those in dashed and dotted lines) form the rest of the network. These edges go through several *hub* processes (which are shown in large fonts representing their high-degree), including DNA recombination, transcription, and DNA dependent regulation of transcription. These are indeed processes that are characterized to mediate genetic regulation. The indirect pathways that go through these mediator processes connect local hubs of the clustered processes, such as response to stimulus and flagellum biogenesis, as well as other fundamental processes including various metabolic and biosynthetic processes, translation, signaling, and transport. These observations illustrate NARADA's ability to accurately capture the basic principles of genetic regulation and characterize the crosstalk between various processes through identification of indirect regulatory relationships.

For the protein interaction network a large portion of the significant pathways involve cellular protein metabolic process, cell cycle, cell division, translation, and response to antibiotic. These hub processes interact with a variety of other biological processes.

A notable problem with projecting networks to the abstract space of functional annotation is that the results are not directly testable. In other words, there is no obvious experimental method that could be used to falsify the notion that a pathway of functional attributes discovered by NARADA is biologicall relevant. These is because, by definition, the pathways identified by NARADA are abstract. Note, however, that patterns identified by NARADA can indeed be used to discover novel biological information that can be experimentally verified, and this provides an indirect method for testing the hypotheses generated by NARADA. Recent applications of frequent pathway templates in Gene Ontology space include functional

annotation of individual proteins [25] and prediction of organism-specific pathways [26].

5. Conclusion

We present a comprehensive software tool, NARADA, to project molecular interaction networks to the functional attribute space. NARADA provides several interfaces to detect significantly overrepresented pathways. Based on results obtained from the *E. coli* transcription network, NARADA identifies several known, as well as novel pathways, at near-interactive query rates. Note that the current knowledge of regulatory networks is incomplete, and is limited to a few model organisms. Therefore, application of our method on currently available data does not 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.

References

1. S. S. Shen-Orr *et al.*, *Nature Genetics* **31**, 64 (2002).
2. M. Koyutürk *et al.*, *Journal of Computational Biology* **13**, 1299 (2006).
3. R. Sharan and T. Ideker, *Nature Biotechnology* **24**, 427 (2006).
4. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).
5. B. Schwikowski *et al.*, *Nat Biotechnol* **18**, 1257(Dec 2000).
6. T. I. Lee *et al.*, *Science* **298**, 799(October 2002).
7. A. H. Y. Tong *et al.*, *Science* **303**, 808(February 2004).
8. J. Gamalielsson *et al.*, A GO-based method for assessing the biological plausibility of regulatory hypotheses, in *ICCS (2)*, 2006.
9. J. Pandey *et al.*, *Bioinformatics* **23**, 377(Jul 2007).
10. S. Liang *et al.*, *Proc. Pacific Symp. Biocomp.* **3**, 18 (1998).
11. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
12. H. Kitano, *Science* **295**, 1662 (2002).
13. V. Chvátal, *Discrete Mathematics* **25**, 285 (1979).
14. P. Shannon *et al.*, *Genome Res* **13**, 2498(Nov 2003).
15. J. Heer *et al.*, Prefuse: a toolkit for interactive information visualization, in *CHI '05: Proceeding of the SIGCHI*, (ACM Press, 2005).
16. H. Salgado *et al.*, *Nucleic Acids Res* **34**(January 2006).
17. L. Salwinski *et al.*, *Nucleic Acids Res* **32**, 449(Jan 2004).
18. E. Camon *et al.*, *Nucleic Acids Res* **32**, 262(Jan 2004).
19. B. Wackwitz *et al.*, *Mol Gen Genet* **262**, 876(Dec 1999).
20. S. Hopper *et al.*, *J Biol Chem* **269**, 19597(Jul 1994).
21. B. Bukau and A. L. Horwich, *Cell* **92**, 351(Feb 1998).
22. C. Chae *et al.*, *J Biol Chem* **279**, 33147(Aug 2004).
23. B. Banecki *et al.*, *Biochim Biophys Acta* **1442**, 39(Oct 1998).
24. K. Saito *et al.*, *J Mol Biol* **235**, 111(Jan 1994), Comparative Study.
25. M. Kirac and G. Ozsoyoglu (submitted).
26. A. Cakmak and G. Ozsoyoglu, *Bioinformatics* (in press).

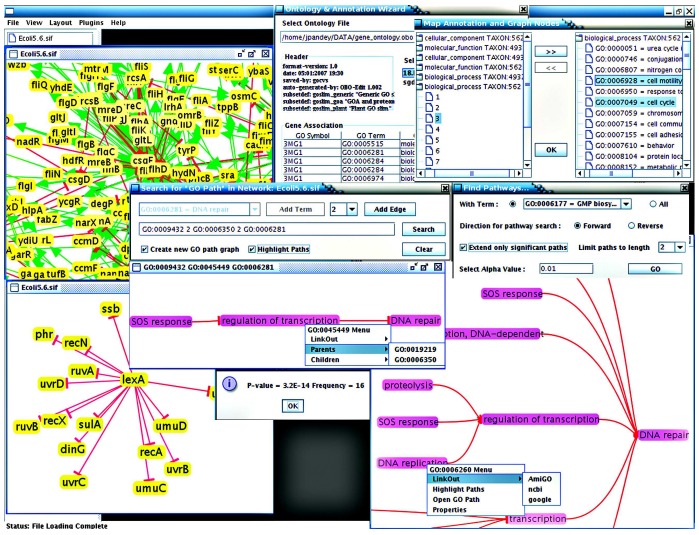
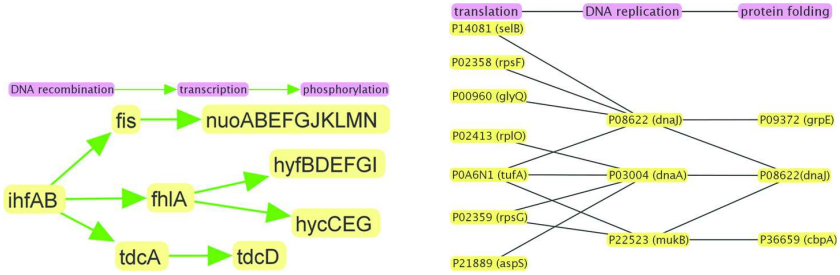


Fig. 3. Screenshots from NARADA.



(a) (b)

Fig. 4. Sample significantly overrepresented pathways in (a) *E.coli* transcriptional network, and (b) *E.coli* protein interaction network. The pathways in functional attribute space are shown on the upper panel, their occurrences in the gene network are shown on the lower panel.

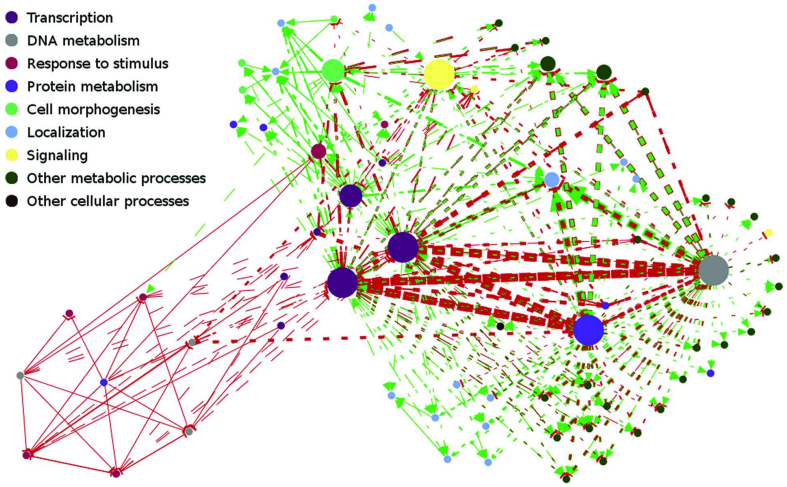


Fig. 5. A global view of *E. coli* transcriptional network mapped to cellular processes defined by GO.