Pacific Symposium on Biocomputing 14:51-62 (2009)

INFERENCE OF FUNCTIONAL NETWORKS OF CONDITION-SPECIFIC RESPONSE - A CASE STUDY OF QUIESCENCE IN YEAST*

SUSHMITA ROY*, TERRAN LANE*, MARGARET WERNER-WASHBURNE‡ AND DIEGO MARTINEZ‡

*Department of Computer Science
‡Department of Biology
1 University of New Mexico,
Albuquerque, NM 87131, USA

Analysis of condition-specific behavior under stressful environmental conditions can provide insight into mechanisms causing different healthy and diseased cellular states. Functional networks (edges representing statistical dependencies) inferred from condition-specific expression data can provide fine-grained, network level information about conserved and specific behavior across different conditions. In this paper, we examine novel microarray compendia measuring gene expression from two unique stationary phase yeast cell populations, quiescent and non-quiescent. We make the following contributions: (a) develop a new algorithm to infer functional networks modeled as undirected probabilistic graphical models, Markov random fields, (b) infer functional networks for quiescent, non-quiescent cells and exponential cells, and (c) compare the inferred networks to identify processes common and different across these cells. We found that both non-quiescent and exponential cells have more gene ontology enrichment than quiescent cells. The exponential cells share more processes with non-quiescent than with quiescent, highlighting the novel and relatively under-studied characteristics of quiescent cells. Analysis of inferred subgraphs identified processes enriched in both quiescent and non-quiescent cells as well processes specific to each cell type. Finally, SNF1, which is crucial for quiescence, occurs exclusively among quiescent network hubs, while non-quiescent network hubs are enriched in human disease causing homologs.

1. Introduction

Cellular adaptations essential for survival under changing enviromental conditions are driven by a complex, but coordinated set of interactions among genes, proteins and metabolites. Existing analyses of condition-specific

*This work is supported by NIMH (1R01MH076282-01, 1R01MH076282-03) NSF (IIS-070568, MCB0734918) and HHMI-NIH/NIBIB (56005678).

Pacific Symposium on Biocomputing 14:51-62 (2009)

behavior typically identify genes differentially expressed across conditions. Fine-grained, interaction analysis among differentially expressed genes can provide deeper insight into human diseases such as cancers[1].

We define functional networks as networks with edges representing general, statistical dependencies among genes. We model functional networks using undirected, probabilistic graphical models, Markov random fields (MRFs). We present a new algorithm, Markov blanket search (MBS), for learning the MRF structure. MBS is based on Abbeel *et al.*'s theoretical work of *Markov blanket canonical parameterization* (MBCP)[2], which requires an exhaustive enumeration ($O(n^l)$) over variable subsets up to size l, where n is the number of variables. We establish an equivalence between the MB canonical parameters and *per-variable* canonical parameters[3], which requires enumeration over only singleton sets ($O(n)$), thus providing a tractable approach for learning genome-scale networks.

We apply our algorithm to two novel yeast (*S. cerevisiae*) microarray datasets measuring gene expression of quiescent and non-quiescent cells, isolated from glucose-starved stationary-phase cultures[4]. Quiescent cells play important roles in health and disease conditions of most living systems, but have been difficult to study due to their low metabolic activity[5]. The recent generation of microarray datasets for these cells[4], gives us the first opportunity to infer a functional network that provides a fine-grained characterization of yeast quiescence.

Algorithms for functional network inference can be broadly classified into: (a) pairwise models (capturing dependencies between only pairs of nodes), and (b) higher-order models (capturing dependencies among two or more nodes). Bayesian networks are higher-order, directed models[6,7,8], but the acyclic constraint of the graph structure cannot easily capture cyclic dependencies. Undirected graphical models can represent cyclic dependencies, but because network inference is much harder[2], higher-order dependencies are often approximated by lower-order (often pairwise) functions[9]. As biological networks are likely to have higher-order dependencies[10], higher-order models are more appropriate for modeling functional networks. Unlike Bayesian networks, our model captures cycles and, unlike pairwise models, we explicitly identify higher-order dependencies.

Bio-techniques for identifying condition-specific networks have been applied to transcriptional regulatory networks[11]. Computational identification of functional networks can provide a less expensive, complementary view of condition-specific networks. Some existing computational approaches integrate mRNA expression with known protein networks[1]. Other

Pacific Symposium on Biocomputing 14:51-62 (2009)

approaches identify cliques of co-regulated genes using mRNA expression. *De novo* functional network inference, can identify novel relationships absent from existing protein networks. Furthermore, identification of general statistical dependencies, including mRNA co-expression, can capture condition-specific responses involving the metabolic and proteomic levels.

We applied MBS to three microarray datasets measuring expression of quiescent, non-quiescent, and exponential yeast cells to genetic and chemical perturbations. We analyze the inferred networks to identify subgraphs that are common among, or that discriminate quiescent, non-quiescent cells, and exponential cells. In both quiescent and non-quiescent cells, fermentation and translation are down regulated, but not in exponential cells. The exponential and non-quiescent cells are more enriched in Gene Ontology[12] (GO) slim processes than quiescent cells, suggesting that the quiescent cells are under-studied novel cell types.

Hub analysis of the inferred networks identified SNF1, which is important for quiescence, to be present exclusively in quiescent cells. Hubs from non-quiescent cells are more enriched in yeast homologs of human disease genes. Overall, our results both agree well with existing knowledge and include novel findings that differentiate quiescent and non-quiescent cells.

2. Methods

2.1. *Markov random fields for biological networks*

A Markov random field (MRF) is an undirected, probabilistic graphical model that represents statistical dependencies among random variables (RVs), $\mathbf{X} = \{X_1, \cdots, X_n\}$. A MRF consists of a graph $\mathcal{G}$ and a set of potential functions $\psi = \{\psi_1, \cdots, \psi_m\}$, which together describe the structural and functional relationships among the RVs. Nodes correspond to continuous RVs encoding gene expression level, $X_i \in \mathcal{R}$.

Although MRFs capture both higher-order and cyclic dependencies, MRF structure learning is harder than in directed models[2]. Approaches to overcome this problem include dependency networks[13] and Markov blanket canonical parameterization (MBCP)[2]. MBCP requires the estimation of optimal Markov blankets (the set of immediate neighbors) for RV subsets, $\mathbf{Y} \subseteq \mathbf{X}$, $|\mathbf{Y}| \leq l$, where l is a pre-specified, maximum subset size.

We extend MBCP by establishing an equivalence between Markov blanket canonical parameters and *per-variable* canonical parameters[3]. As a consequence, we need to estimate Markov blankets of individual RVs instead of all subsets. Abbeel *et al*'s MBCP exhaustively enumerates over subsets

Pacific Symposium on Biocomputing 14:51-62 (2009)

up to size l taking $O(n^l)$ time. Instead, our approach enumerates over singleton sets resulting in a reduced complexity of $O(n)$, producing a more tractable structure learning approach.

Our algorithm also enforces structural consistency, which is not guaranteed in dependency networks and MBCP. In structurally consistent networks, for every pair $\{X_i,X_j\}$, X_i is in X_j's Markov blanket, if and only if, X_j is in X_i's Markov blanket. A structurally inconsistent network may not have an associated joint probability distribution making probabilistic inference difficult.

2.2. *Markov blanket search algorithm*

Our per-variable canonical parameterization enables structure search by identifying the best Markov blanket (MB) per RV[3]. We identify the best MB by minimizing conditional entropy[14,2], $H(X_i|\mathbf{M}_i)$ for each X_i given its MB, $\mathbf{M}_i$. We enforce structural consistency by computing the score gain on adding an edge, $\{X_i, X_j\}$. This combines the decrease in $H(X_i|\mathbf{M}_i)$ due to addition of X_j, with the conditional entropy change if X_j's MB was constrained to include X_i. Overall, MBS minimizes $\sum_{i=1}^{n} H(X_i|M_i)$ plus a regularization term, subject to the structure consistency constraint.

Let $\mathbf{M}_i^k$ denote the current best MB for X_i of size k, X_j denote a candidate for addition to $\mathbf{M}_i^k$, and $\mathbf{M}_j^k$ be the current best MB for X_j . Then the score gain is:

$$G_i = H(X_i|\mathbf{M}_i^k) - H(X_i|\mathbf{M}_i^k \cup \{X_j\}) + H(X_j|\mathbf{M}_j^k) - H(X_j|\mathbf{M}_j^k \cup \{X_i\}). \quad (1)$$

The MBS algorithm performs a greedy search to identify the best MB for each variable. Each search iteration uses a combination of *add* and *swap* operations. In the add stage of the $k+1^{th}$ iteration, we make one variable extensions to the current Markov blanket $\mathbf{M}_i^k$ of each X_i restricting it to at most $k+1$ RVs per MB. In the swap stage, we revisit all variables Z in the Markov blanket $\mathbf{M}_i^k$ of each X_i, and consider other RVs $Y \notin (\{X_i\} \cup \mathbf{M}_i^k)$, which if swapped in instead of Z, gives a score improvement. We assume all variables to have a Gaussian distribution. However, our general approach is applicable to other random variables, requiring only the estimation of conditional entropy.

2.3. *Data pre-processing*

We applied our algorithm to two yeast, *S. cerevisiae*, datasets from quiescent and non-quiescent cells[4], and one dataset from exponential cells[15]. We

included only genes with $< 20\%$ missing data in all three datasets. As the quiescent and non-quiescent datasets had biological replicates, we filtered the genes further to discard not-reproducible genes. Our final datasets comprised $n = 2,818$ genes, with 170, 186, and, 300 measurements per gene in the quiescent, non-quiescent and exponential populations respectively.

2.4. *Analysis of inferred networks*

Generation of coarse modules: To obtain a high-level view of the inferred networks, we generated local subgraphs and clustered them into coarse modules. We excluded clusters of size < 5 and connected the remaining clusters into high-level graphs (Fig 1, Section 3.1).

We generated subgraphs by considering a node, and its neighbors reachable by r links. We refer to these subgraphs as 1-n subgraphs, denoting a $n$eighborhood reachable by traversing one link ($r = 1$). We computed a topological similarity for each pair of subgraphs, $\{S_i, S_j\}$, $t_{ij} = \frac{l_{ij}}{n_i+n_j}$, where l_{ij} is the sum of number of vertices common between S_i and S_j, and the number of edges across S_i and S_j. n_i and n_j are the vertex counts in subgraphs, S_i and S_j, respectively[16].

To obtain *coarse modular organization*, we first clustered the subgraphs using hierarchical clustering with average linkage[17]. We selected clusters to optimize between including majority of the genes, and to have clusters of size ≥ 5. This resulted in 230, 214 and 179 clusters, with $n = 2630, 2551$ and 2651 genes in quiescent, non-quiescent and exponential cells respectively. We then used topological similarity as edge weights for each pair of clusters.

To assess if the clusters were biologically meaningful, we computed an *annotation similarity*, f_{ij}, for each cluster pair, $\{C_i, C_j\}$. We obtained GO slim process enrichment vector, e_i per cluster. Each dimension $e_i(r)$ was the logarithm of p-value enrichment for each process term. The annotation similarity between C_i and C_j was the Pearson's correlation coefficient between e_i and e_j.

We developed a measure, *Annotation-Topological Similarity* (ATS), to assess if clusters that were topologically close were also similarly annotated. ATS is the Pearson's correlation coefficient between two vectors, v_A and v_T, each of length $\binom{|\mathbf{C}_x|}{2}$, where $\mathbf{C}_x$ is the set of clusters generated for population x. Each dimension $v_A(r)$ ($v_T(r)$) was the annotation (topological) similarity for r^{th} cluster pair, where $1 \leq r \leq \binom{|\mathbf{C}_x|}{2}$.

We define *relative enrichment* among two populations x and y that tests if x is equally, less or more annotated than y. Let p_y be the proportion of

Pacific Symposium on Biocomputing 14:51-62 (2009)

y's clusters that are enriched ($< 10^{-2}$) in any slim term. Assuming r out of s clusters of x are enriched, we compute the probability of observing $\leq r$ out of s using the binomial with parameter p_y. The smaller the probability the more depleted is x as compared to y. Similarly, the probability of $\geq r$ out of s enriched clusters estimates how annotated x is as compared to y. Smaller the probability the more annotated x is w.r.t to y. We repeat this for testing y's relative enrichment w.r.t x. We do this analysis for all population pairs such as quiescent versus exponential, quiescent versus non-quiescent, etc.

Identification of up or down-regulated subgraphs: We analyzed the 1-n subgraphs for *significant up-regulation* or *down-regulation* at expression level using a similar approach to Chuang *et al.*[1] Each subgraph was assigned the average of the mean expression of the subgraph genes. For each dataset and subgraph size, we estimated a null distribution of subgraph expression by randomly sampling $s = 100,000$ subsets of all genes. A p-value < 0.05 was considered as significantly up or down-regulated.

Identification of conserved and specific subgraphs: To identify *conserved subgraphs* among two cell populations, A and B, we computed a match score for each 1-n subgraph, $S_i^A \in \mathbf{S}_A$ generated from A's network, using B's network structure. This score is the harmonic mean of recall, R_i^A, and precision, P_i^A, per subgraph S_i^A. For each $S_i^A \in \mathbf{S}_A$, $R_i^A = \frac{|E_i^A \cap E_i^B|}{|E_i^A|}$, where E_i^A is the edgeset of S_i^A in A's network, and E_i^B is the edge set among S_i^A's vertices in B's network. Similarly, $P_i^A = \frac{|E_i^A \cap E_i^B|}{|E_i^B|}$. The match of S_i^A in B's network is $F_i^A = \frac{2 * P_i^A * R_i^A}{P_i^A + R_i^A}$. We assumed an $F_i^A > 0$ to indicate S_i^A is a conserved subgraph in B's network. The set of conserved subgraphs between A and B is $\mathbf{S}_A^M \cup \mathbf{S}_B^M$, where $\mathbf{S}_A^M \subseteq \mathbf{S}_A$ and $\mathbf{S}_A^M \subseteq \mathbf{S}_B$. Each $S_i^A \in \mathbf{S}_A^M$ has $F_i^A > 0$, when compared with B's network, and each $S_j^B \in \mathbf{S}_B^M$ has $F_j^B > 0$, when compared with A's network. A subgraph was considered *specific* to a particular population if it had a match score of zero for the remaining populations.

Gene ontology (GO) enrichment and false discovery rate: For each 1-n subgraph (or cluster), we used the hyper-geometric distribution to compute GO term enrichment. We sampled $s = 1000$ random subsets of size k from the $n = 2818$ genes, and computed their p-values for each term. The false discovery rate (FDR)[18] is associated with the number of terms enriched in a subgraph at a particular p-value. For example, if a subgraph of size k is enriched in u terms at $p < 10^{-4}$, FDR is $\frac{u'}{u}$, where u' is the average

number of terms enriched in a random subgraph of size k at $p < 10^{-4}$. We used an FDR of ≤ 0.05 to identify significant enrichments.

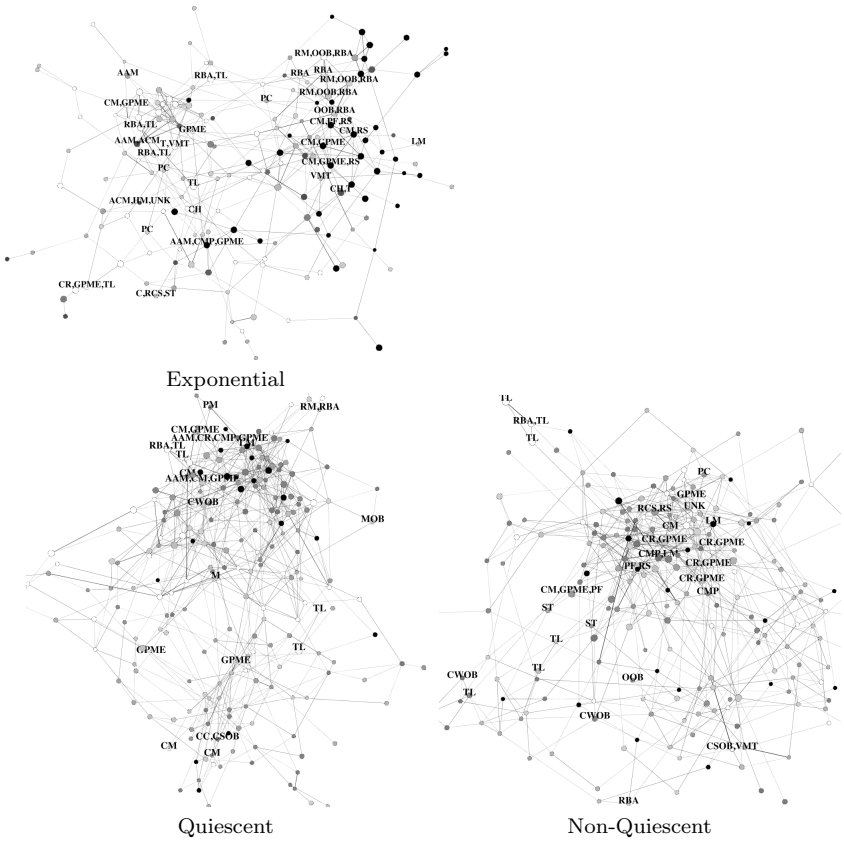

Figure 1. Coarse modular organization of networks inferred from three populations. Each node represents a cluster of genes. The node size is proportional to the gene count per cluster. The node shade indicates if cluster genes are more expressed (dark) or repressed (light). The edge thickness is proportional to similarity between two clusters. The nodes are labeled with processes enriched in the member genes. AAM: amino acid and deriv. ACM: aromatic comp CM: carb. meta CC: cell cycle CH: cellular homeostasis CR: cellular resp. CWOB: cell wall org. & biogen. CMP: cofactor metab. C: conjugation CSOB: cytoskeleton org. & biogen. GPME: gen. of pre. metab. &energy HM: heterocycle metab. LM: lipid metab. M: meiosis MOB: membr. org. & and biogen. OOB: organelle org & biogen. UNK &other PC: protein catab. PF: protein folding PM: protein modif RCS: resp. to chemical stimulus RS: response to stress RBA: ribosome biogen & assem RM: RNA metab ST: signal transduction TR: transcription TL: translation T: transport TP: transposition VMT: vesicle-mediated transport VM: vitamin metabolic process.

Pacific Symposium on Biocomputing 14:51-62 (2009)

3. Results

3.1. *Modular organization in quiescent, non-quiescent and exponential cells.*

To obtain a high-level view of inferred networks, we clustered the subgraphs from inferred networks per population, followed by GO slim process enrichment of the clusters (Fig 1). We had high Annotation-Topological similarity (ATS) measure (see methods), for exponential (ATS=0.51) and non-quiescent cells (ATS=0.49), suggesting that similarly functioning genes are topologically close in the inferred networks. Quiescent cells had lower agreement (ATS=0.26), suggesting that gene expression is more informative for non-quiescent and exponential cells, than quiescent cells. Quiescent cells may employ additional mechanisms, including post-translational modification, to respond to stresses[19,4]. The presence of post-translational modification is further supported by up-regulation of a quiescent cluster enriched in protein modification (PM).

Relative enrichment analysis shows that both quiescent and non-quiescent cells have significantly fewer annotated clusters than exponential (Table 1). Quiescent is also less annotated than non-quiescent. This highlights the under-studied characteristics of quiescent cells, motivating further investigation of these cells.

Table 1. Relative enrichment of clusters. ↑ and ↓ denote enrichment and depletion p-value, respectively, of annotated clusters.

Population	Enriched/Total clusters	wrt EXP		wrt Q		wrt NQ	
		↑	↓	↑	↓	↑	↓
Exponential	32/179	–	–	3e-4	1.0	0.03	0.94
Quiescent	19/230	1	7e-5	–	–	1.0	0.02
Non-quiescent	27/214	0.04	0.96	0.01	0.99	–	–

3.2. *Fine grained analysis of the cell populations*

3.2.1. *Quiescent and non-quiescent cells show common global starvation response behavior.*

To identify similarly enriched processes, we obtained conserved subgraphs among two populations. For each subgraph, we compared GO process enrichment, and whether it agreed in expression – both up or both down-regulated – in the two populations being compared.

There were a large number of subgraphs common between quiescent and non-quiescent (Table 2). These subgraphs were enriched in glycolysis,

Pacific Symposium on Biocomputing 14:51-62 (2009)

Table 2. Number of conserved subgraphs between pairs of populations. Enr implies enriched. Up and Down indicate significant up and down regulation, respectively.

Population pair	Subgraphs	Enr	Up	Down	Enr & Up	Enr & Down
Q-NQ	833	97	26	95	4	27
NQ-EXP	288	99	7	25	5	16
Q-EXP	311	98	0	27	11	24

fermentation, translation and fatty acid oxidation processes. However, only half agreed in expression. Several of these subgraphs had both up and down-regulated genes, resulting in only average expression of the entire subgraph. These heterogeneous dependencies indicate more complex relationships not likely to be captured by co-expression.

We also identified several subgraphs conserved between quiescent and non-quiescent populations, that were up-regulated, but did not have term enrichment. Genes from these subgraphs were associated with unknown biological process, further elucidating the importance of studying these cells.

Non-quiescent and exponential cells had several conserved subgraphs enriched in telomere maintenance, DNA packaging, chromatin assembly and mitotic recombination. These findings are consistent with previous knowledge of non-quiescent cells to have unstable genomes, and, therefore requiring these processes[5]. Comparison of quiescent and exponential cells did not identify any processes enriched in the up-regulated subgraphs. The processes enriched in down-regulated subgraphs included glycolysis, gluconeogenesis and ribosomal biogenesis.

Overall, the subgraph analysis suggests that quiescent and non-quiescent cells are more similar to each other, than each is to exponential cells. There are several subgraphs common to quiescent and non-quiescent cells, but not all agree in expression. The processes that are common between these cells suggest global environmental response as the cells transition from fermentable to non-fermentable carbon sources for energy.

Table 3. Subgraphs specific to individual populations. Same legend as Table 2.

Population	Subgraphs	Enr	Up	Down	Enr & Up	Enr & Down
Q	2295	70	174	160	7	17
NQ	2317	74	171	186	10	11
EXP	2570	232	327	206	54	44

3.2.2. *Differences in quiescent and non-quiescent cells suggest population-specific response*

We examined GO enrichment of subgraphs that occurred only in one population. Both quiescent and non-quiescent cells had fewer subgraphs with

Table 4. Processes exclusively up regulated in different populations

Population	Process	genes
Q	RAS signal transduction	IRA1, SPG3, YGR026W, BCY1, PFK2
	Sporulation	GPA2, GSC2, OSW2, CAF120, YOR277C
	de novo pyrimidine base biosynthetic process	ARF1,DIG1, URA1,URA3,YHR003C
NQ	hyperosmostic response	TRS120, MSB2, YHR100C, PBS2, RSF2
	regulation of DNA metabolic process	PIM1, DIG2, BCK2,SSL2, DPB11, PLB3, HST1, YNG1
EXP	ATP biosynthesis	COX20, QCR10,QCR8, ATP2, ATP7
	cell wall organization & biogenesis	AFR1, SKN1,GFA1, KTR2, DFG5
	amino acid biosynthesis	IDP1, ARO3, HOM2, YGL117W, YSC83, ARG4,SIP4, CPA2,ARG1, SER1, SSU1
	response to toxin	AAD10,AAD16, AAD4,BAP2,MID2, TAT1,TYR1

enriched processes than exponential (Table 3). The quiescent cells were exclusively enriched in sporulation and negative regulation of the RAS signal transduction pathway (Table 4). Down regulation of this pro-growth pathway indicates mechanisms to conserve energy expended in growth conditions. Furthermore, subgraph genes that are not annotated with signal transduction (SPG3, PFK2), are all important for stationary phase.

The non-quiescent cells exhibited processes involved in osmotic stress response and regulation of DNA recombination. This is consistent with these cells trying to cope with environmental changes and that they have unstable genomes. However, unlike quiescent cells, most of the processes up-regulated in non-quiescent cells, also occurred in exponential cells.

The exponential cells were enriched in response to chemical stresses, biosynthesis of amino acids and ATP biosynthesis. ATP biosynthesis was down-regulated in both quiescent and non-quiescent cells. The up-regulation of these energy producing pathways suggests that exponential cells expend a large amount of energy to make relevant mRNA in response to different stresses. In contrast, as quiescent cells are formed in response to a starvation condition, they are likely to sequester mRNA for rapid release in response to different stresses[19].

3.2.3. *Non-quiescent hubs are enriched in disease causing genes*

We analyzed the inferred networks to identify network hubs, nodes with degree ≥ 7. A significant overlap between quiescent and non-quiescent hubs (n=29) implied similarities among these cells due to global starvation response, consistent with Section 3.2.1.

Pacific Symposium on Biocomputing 14:51-62 (2009)

Table 5. Hub nodes and their most enriched processes

Population	Hubs	Exclusive	Processes
Q	215	167	cell wall organization & biogenesis signal transduction, carbohydrate metabolism, organelle organization and biogenesis, generation of precursor metabolites
NQ	166	116	vesicle-mediated transport, response to stress, membrane organization & biogenesis
EXP	318	273	aminoacid & derivative process, cellular respiration ribosome biogenesis & assembly

The non-quiescent cells have been hypothesized as models for studying diseases in humans due to the instability of their genomes[4]. We asked if hubs from different cell populations were enriched in human disease causing gene homologs[20] (Table 6). Of the $n = 2818$ genes used to infer networks, there were $n = 225$ yeast genes, homologous to different human disease genes[a]. We found that hubs in non-quiescent cells are more likely to be enriched in disease homologs than either quiescent or exponential cells. This provides preliminary empirical evidence for the hypothesis that these cells can provide insight into human disease causing conditions.

We found network hubs from quiescent cells to be enriched (p-value <0.05) in signal transduction and cell wall biogenesis (Table 5). Among the quiescent hubs was SNF1, known to be crucial for the formation of quiescent cells. The non-quiescent hubs were enriched in stress response and vesicle mediated transport. Finally the exponential hubs were enriched in amino acid processes and cellular respiration. The enrichment of different processes further illustrates the underlying bio-chemical characteristics that discriminate these cells, and how they respond to different stresses.

Table 6. Enrichment of human disease gene homologs in hubs.

Population	Total Hubs	Homologous Disease Hubs	Pval
NQ	166	26	5e-4
Q	215	22	0.130
EXP	318	26	0.395

4. Conclusion

We have developed an undirected, probabilistic graph learning algorithm that can capture different types of dependencies that may exist in expression-based, functional networks. We use our algorithm to perform

[a]We downloaded human-yeast homologs from `http://www.biomart.org/index.html`

a network analysis of three yeast cell populations in starvation and exponential growth conditions. A high-level analysis of the modular structure of these networks suggests that quiescent cells are significantly underannotated, highlighting the need to study these cells. Our analysis suggests that the non-quiescent cells share more characteristics with exponential cells as compared to quiescent. Analysis of individual subgraphs indicates that quiescent and non-quiescent cells exhibit similarities in their mechanisms to adapt to glucose starvation. However, there are processes specific to quiescent cells such as sporulation, which suggest alternative response mechanisms that might be active in these cells. Finally, we find that non-quiescent hubs are enriched in homologs of human disease genes. In summary, our network-based analysis has identified both previously known and novel biological processes that are important in these cells, giving a finer understanding of the mechanisms conserved and specific to these cells.

References

1. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, *Mol Syst Biol* **3** (2007).
2. P. Abbeel, D. Koller, A. Y. Ng, *JMLR* **7**, 1743 (2006).
3. S. Roy, T. Lane, M. Werner-Washburne, *Tech. Rep. TR-CS-2008-14* , University of New Mexico (2008).
4. A. D. Aragon, *et al.*, *Molecular Biology of the Cell* (2008).
5. C. Allen, *et al.*, *J Cell Biol* **174**, 89 (2006).
6. N. Friedman, *Science* **303**, 799 (2004).
7. J. Yu, A. A. Smith, P. P. Wang, A. J. Hartemink, *Bioinformatics* **20**, 3594+ (2004).
8. E. Segal, *et al.*, *Nat Genet* **34**, 166 (2003).
9. A. Margolin, *et al.*, *BMC Bioinformatics* **(Suppl 1): S7** (2005).
10. Y. Qi, H. Ge, *PLoS Computational Biology* **2**, e174+ (2006).
11. C. T. Harbison, *et al.*, *Nature* (2004).
12. M. Ashburner, *et al.*, *Nat Genet* **25**, 25 (2000).
13. D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, C. M. Kadie, *JMLR* **1**, 49 (2000).
14. T. M. Cover, J. A. Thomas, *Elements of information theory* (Wiley-Interscience, New York, NY, USA, 1991).
15. T. R. Hughes, *et al.*, *Cell* **102**, 109 (2000).
16. I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, *Science* **306**, 1555 (2004).
17. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (1998).
18. E. I. Boyle, *et al.*, *Bioinformatics* **20**, 3710+ (2005).
19. A. D. Aragon, G. A. Quiñones, E. V. Thomas, S. Roy, M. Werner-Washburne, *Genome Biol* **7** (2006).
20. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, *Nucleic Acids Res* **33 Database Issue** (2005).