BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i354–i358
doi:10.1093/bioinformatics/btm172

Genomic characterization of perturbation sensitivity

Jung Hun Ohn[1], Jihun Kim[1] and Ju Han Kim[1,2,*]

[1]Seoul National University Biomedical Informatics (SNUBI) and [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

ABSTRACT

Motivation: In determining the function of a gene, it provides much information to observe the changes in a biological system after disruption of the gene of interest through its knockout. Thanks to the microarray technology, it is now possible to profile transcriptional changes of the whole genome, thus differentiating genes that are significantly affected by the knockout. Based on microarray experiments of hundreds of different knockouts, we assigned the so called, 'Perturbation Sensitivity', to the *Saccharomyces cerevisiae* genome by the frequency of significant changes in the transcript level in hundreds of knockout conditions. Biologically, it reflects the degree of a gene's sensitivity to external perturbations.

Results: Through gradually enriching gene sets with more perturbation sensitive genes, we show that perturbation sensitive genes are usually not essential and their coding proteins have fewer physical interaction partners and more transcription factors bind to their upstream sequences. And the two extreme gene groups, perturbation sensitive versus perturbation resistant, have mutually exclusive functional annotations.

Contact: juhan@snu.ac.kr

1 INTRODUCTION

We perturb a system to understand it. In biology, one of the most widely employed approaches to understand the function of a gene is to completely or partially disrupt the gene of interest and observe phenotypic changes in the transformed animal (knockout mice) or the cell (deletion mutants). With the advent of high-throughput technologies like DNA microarray the transcriptional activity of thousands of genes are measured simultaneously and it is now possible to observe not just phenotypic changes but also the ups and downs of thousands of genes in response to gene disruptions. We can reliably assume the existence of direct or indirect transcriptional relationship between the disrupted gene and the significantly up or down regulated gene secondary to the disruption.

This cause and effect relationships between genes (the disruption of gene *A* leading to the transcriptional change of gene *B*) were modeled in the Bayesian Network approaches (Friedman, 2004; Pe'er *et al*., 2001). From the transcriptional profiling study of the yeast genome in response to 276 different gene disruptions (Hughes *et al*., 2000), the approach aims to extract graph structures that represent statistical dependency and causal relationship among genes. But the extracted graph structure deals with only subsets of genes out of thousands of yeast genes (Markowetz *et al*., 2005).

On the other hand, the 'disruption network' approach (Featherstone and Broadie, 2002; Rung *et al*., 2002; Wagner and Fell, 2001) is simpler but genome-wide. It is defined as a directed graph where nodes represent genes and arcs link nodes if the disruption of the source gene significantly alters the target gene. They explored the disruption network for network properties like degree centrality, scale-free topology and modularity.

Now, rather than extracting biologically meaningful graph structures within each network, we address an important question which is only possible with such genome-wide transcriptional profiling study of large scale gene disruptions. If there are genes vulnerable or resistant to have changes in transcription levels after perturbations to a system, i.e. gene disruptions, what are their genomic characteristics and what are the biological insights? Here, the genomic characteristics refer to phenotypic information, topological position in protein-to-protein interaction network or transcriptional regulatory network.

From the above mentioned transcriptional profiling study of the yeast genome in response to 276 different gene disruptions (Hughes *et al*., 2002), the 'perturbation network' is defined as a *non-directed bipartite* graph of two node groups; a group of 'genes' which showed significant changes in transcription level in the other group of 'deletion mutants' and links are made between nodes from each group based on the significance level assuming the 'error model'. This is a non-directed version of the above mentioned 'disruption network' and is made up of 4280 genes and 212 deletion mutants.

We sorted the 4280 genes according to its degree whose biological meaning is straightforward; how sensitive is the gene to external perturbations? Therefore we termed the degree as 'perturbation sensitivity' of the gene. From the network, we sliced out the genes with low sensitivity, thus enriching the gene set with gradually more perturbation sensitive genes to find that perturbation sensitive genes are usually not essential and their coding proteins have fewer physical interaction partners and more transcription factors bind to their upstream sequences. Further, it is also explored what functional categories are significantly overrepresented in each perturbation sensitive and resistant gene set.

2 METHODS

2.1 Definition of 'perturbation network'

Perturbation network is constructed from the genome-wide transcriptional profiling study of 300 perturbation experiments like gene deletions or drug treatments in *Saccharomyces cerevisiae* (Hughes *et al*., 2000). The dataset includes mRNA expression profiles of 6325 yeast ORFs in 276 different single-gene deletion mutant strains

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

and is available at http://rii.com/tech/pubs/cell_hughes.htm. We investigated only gene deletion experiments because drug treatments usually perturb more than one gene and could increase heterogeneity in the dataset.

A graph is *bipartite* if the vertices are partitioned in two mutually exclusive sets such that there are no ties within either set and every edge in the graph is an unordered pair of nodes in which one node is in one vertex set and the other in the other vertex set. (Wasserman and Faust, 1994).

Perturbation network is a *non-directed bipartite* graph where one vertex set contains genes significantly up or down regulated in deletion mutants constituting the other vertex set and a link is made between gene i and deletion mutant j if the expression of gene i is significantly altered in deletion mutant j.

Based on the 'error model' (Hughes *et al.*, 2000) correcting for gene measurement error and biological noise, P-value is assigned for each pair of gene and deletion mutant. A total of 4280 genes showed any significant changes in more than one deletion mutants and 212 deletion mutants affected more than one gene according to the criteria: P-value < 0.01. The original set of 276 deletion mutated genes reflects a variety of functional categories according to the MIPS (Munich Information Center for Protein Sequence) functional catalogues (Hughes *et al.*, 2000; Mewes *et al.*, 2002) and the selected 212 deletion mutated genes had similar distribution in functional catalogues, excluding possible bias from the cut-off process (Data not shown).

In matrix notation, perturbation network is represented as matrix $D = \langle d_{ij} \rangle$ for gene i and deletion mutant j where,

$$d_{ij} = \begin{cases} 1, & P < 0.01 \\ 0, & P \geq 0.01 \end{cases}$$

2.2 Perturbation sensitivity (S_i)

For each gene i, the perturbation sensitivity (S_i) is defined as its node degree in the bipartite graph.

$$S_i = \sum_i \mathrm{d_{ij}}$$

which is the number of deletion mutants in which the gene is differentially expressed. The larger S_i value means that gene i is up- or down-regulated in a larger number of deletion mutants and highly sensitive and responsive to external perturbations. In the current dataset, it ranged from 1 to 49.

2.3 Slicing gene group by perturbation sensitivity

We sorted the 4280 genes according to its S_i. Then genes with the least S_i's are shaved out to enrich the group with genes with higher S_i's. Here we define a gene group with S_i's equal to or higher than m as *m-core*. After slicing out genes with S_i's equal to m, $(m+1)$-*core* remains and is nested in *m-core*. Iteratively applying the procedure produces groups of gradually more perturbation sensitive genes.

2.4 Excess retention

Excess retention (Wuchty and Almaas, 2005) is defined as the degree to which genes with a certain property A is over or underrepresented in *m-core* compared to that in the whole gene group or *1-core*. The fraction of genes with property A in the whole group of N genes is $E^A = N^A/N$. If *m-core* contains N_m genes and the number of genes with property A in *m-core* is N_m^A, then the excess retention of genes with property A in *m-core* is given by

$$E_m^A = \frac{(N_m^A/N_m)}{(N^A/N)}.$$

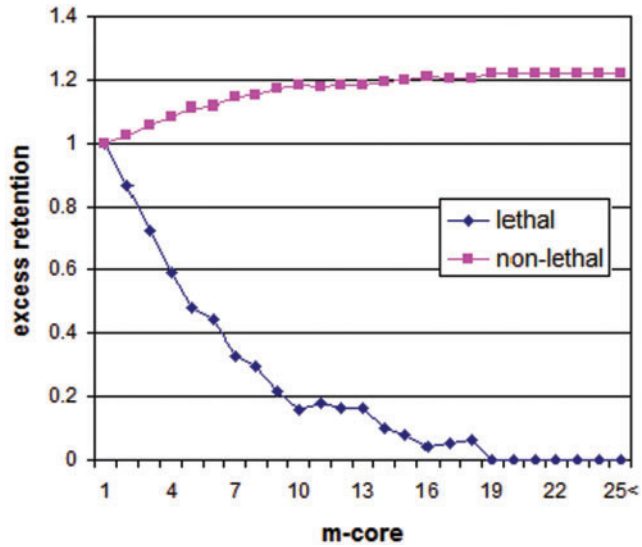

Fig. 1. The excess retention of non-lethal genes among perturbation sensitive genes.

3 RESULTS

The perturbation sensitivity is a novel transcriptomic property. Previous genome-wide studies have characterized a gene with various genomic properties like the impact on the viability of a cell in its absence (Winzeler *et al.*, 1999) and topological position in protein interaction (Uetz *et al.*, 2000) or transcriptional network (Lee *et al.*, 2002).

To uncover the implication of a gene's transcriptional sensitivity to external perturbations, we first investigate the associations between the 'perturbation sensitivity' and the previously well studied genomic characteristics.

3.1 The enrichment of non-lethal genes among perturbation sensitive genes

At the most basic level, the functional importance of a gene is defined by its lethality or essentiality (Winzeler *et al.*, 1999). A gene is usually called lethal (in other words, non-viable or essential) if its absence leads to death of its deletion mutant and non-lethal (in other words, viable or not essential) if its absence does not affect the survival of the cell.

The *Saccharomyces* Genome Database (http://www.yeastgenome.org) provides the lethality information of yeast genes. Of the 4280 genes, 696 genes were lethal and 3214 genes were non-lethal and the rest had no lethality information. We studied the excess retention (see *methods*) of lethal and non-lethal genes in each *m-core*. If the perturbation sensitivity and lethality are not correlated, the ratio of the lethal or non-lethal genes in each *m-core* to the expected number would be around one. Instead, Figure 1 shows the excess retention of non-lethal genes (with peak of 1.2-fold) while lethal genes are diluted to 0 in cores >19.

In a sense, lethal and essential genes might be more likely to be up- or down-regulated because they participate in more biological processes and have a longer evolutionary history (Jeong *et al.*, 2001). But, in our results, the accumulation of

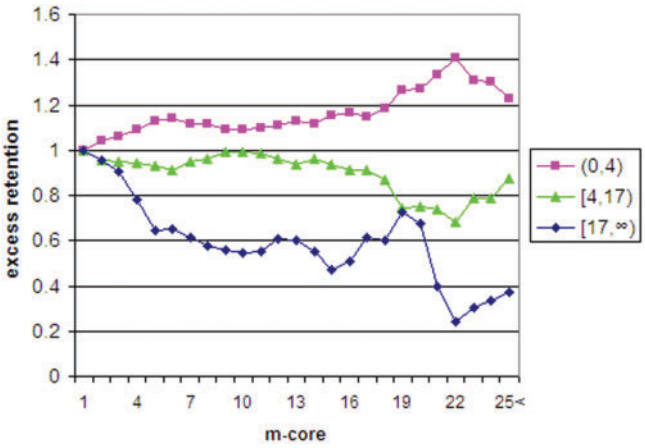

Fig. 2. The excess retention of genes with less physically interacting partners among perturbation sensitive genes.

non-lethal genes in the higher cores suggests that non-lethal genes are more liable to show transcriptional changes in numerous random perturbations on the system while lethal and essential genes are robust to random perturbations and tolerant to outside attacks. Despite vigorous interventions like gene disruptions, genes essential for maintaining the life of the organism are not easily up- or down-regulated. This accords with the earlier finding by Albert *et al*. that the biological network shows a surprising degree of tolerance against errors which is attributable to its scale-free nature (Albert *et al*., 2000).

3.2 The depletion of physical interaction network hubs among perturbation sensitive genes

The tremendous accumulation of protein interaction data has made it possible to construct the protein interaction networks and the network topology has been widely studied (Uetz *et al*., 2000). Now the topological position of the proteins coded by the perturbation sensitive genes in the protein interaction network is explored using the 'excess retention' approach. The perturbation sensitivity is based on the behavior of individual gene in *transcriptional* level and how the property is related to the molecular interactions in *proteome* level is a very interesting issue.

We retrieved from the *Saccharomycess* Genome Database (http://www.yeastgenome.org) physical interaction data of yeast genes. A total of 3736 genes had more than one physically interacting partners and the number of interactions varied from 1 to 619. It is categorized into three levels or low, intermediate and high. Two cut-off values, 4 and 17 are arbitrarily chosen to represent half and upper 15 percentile, respectively.

Figure 2 shows the excess retention of the three categories in each *m-core*. Perturbation sensitive gene group is enriched in genes whose coding proteins have less physical interactions while proteins coded by genes that show stable expressions against perturbations have more physically interacting partners.

Intuitively, if a gene transcript or a protein is interacting with many other proteins, it is more likely to be affected by external perturbations because it is linked to many cellular processes. Instead, our result suggests these physical interaction network hub-proteins or proteins with large number of partners, show high stability in transcription levels to such interventions. They are placed in the periphery in perturbation network and this well demonstrates that physical interaction and transcriptional changes are definitely two different layers of cellular processes.

In Section 3.1, we showed that perturbation sensitive genes are usually non-lethal. The observations that highly interacting proteins have an increased tendency to be lethal (Jeong *et al*., 2001), coupled with the association between the lethality and perturbation sensitivity in Section 3.1, make it possible to presume the result in Figure 2 indirectly. Here, we establish the direct link between the perturbation sensitivity and protein interaction topology.

In biological sense, physical interaction network hubs which are usually essential (Jeong *et al*., 2001) might be the 'house keeping' genes, as suggested by Yu *et al*. (Yu *et al*., 2004). They are constitutively transcribed and are very resistant to external perturbations through some kinds of buffering systems but cause deleterious impact on the individual when affected by a perturbation.

3.3 Perturbation sensitive genes have more transcriptional regulators

Recently, transcriptional regulatory networks are constructed and explored through genome-wide profiling studies like ChIP-on-Chip, which uses chromatin immunoprecipitation and DNA microarrays together with computational analysis to map the genomic sites bound by transcription factors (Lee *et al*., 2002). Such technologies have made it possible to study the relationship between transcription factors and downstream genes in a quantitative way.

Perturbation sensitive genes are, by definition, differentially transcribed in a variety of environmental challenges, i.e. gene knockout conditions. Now we address the question if the observed transcriptional responses are correlated with the number of transcription factors binding to upstream sequence of genes.

The YEASTRACT database ('Yeast Search for Transcriptional Regulators And Consensus Tracking', http://www.yeastract.com) incorporates data from genome-wide technologies and literatures. It is a web-based service providing a list of transcription factors for a group of genes and we used 4280 genes as input to get 3017 genes with information on regulating transcription factors. The number of transcriptional regulators varied from 1 to 22 and was categorized into three groups of genes; low, intermediate and high using three (50 percentile) and seven (upper 10 percentile) as cut-off values.

As shown in Figure 3, perturbation sensitive genes are controlled by *more* transcription factors (maximum of more than 3-fold than expected in 'high' group). 'Intermediate' group showed only mild retention with higher *m*. We come to the conclusion that perturbation sensitivity is *quantitatively* correlated with the number of upstream transcriptional regulatory factors.

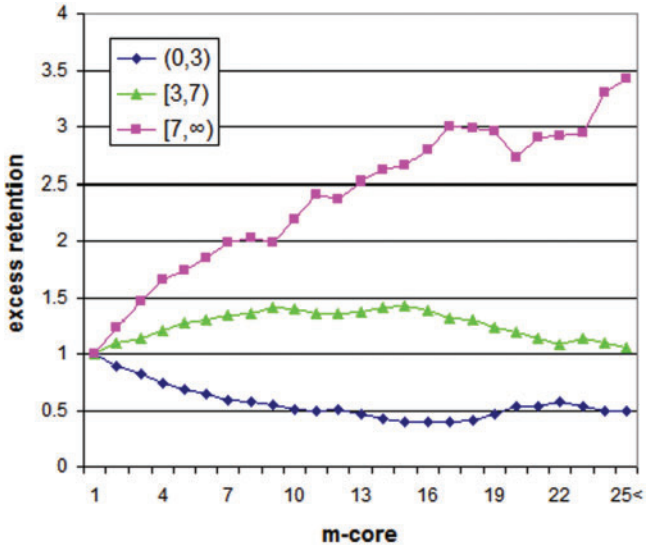

Fig. 3. The excess retention of genes with more transcription factors bound to upstream sequence among perturbation sensitive genes.

3.4 The enrichment of functional categories

In addition to exploring correlations among objective properties of genes, we also investigated if the group of perturbation sensitive and insensitive genes has significant proportion of genes with certain functional annotations. The MIPS database (http://mips.gsf.de) contains functional catalogues which is a functional annotation scheme for systematic classification of proteins from whole genomes (Mewes *et al.*, 2002). And it also provides on-line tools (http://mips.gsf.de/proj/funcatDB) for statistical test of significant enrichment of a given gene set in certain functional categories as compared to the set of whole genome under the assumption of the hypergeometric distribution (Ruepp *et al.*, 2004).

Figure 4 gives the distribution of significantly overrepresented functional categories (P-value < 0.01) among perturbation sensitive genes (genes nested in the *6-core* which is made up of 831 genes). The most highly perturbation sensitive group is enriched in genes that belong to the two functional categories, *metabolism* and *interaction with the cellular environment*. *Metabolism* associated genes are highly enriched throughout *m-cores* and it is natural to suppose that genes participating in the *interaction with the cellular environment* are perturbation sensitive as they should be actively transcribed or repressed according to changing cellular environments like the deletion of related genes.

To represent a perturbation resistant or insensitive group, 831 genes, the same size of the *6-core*, were randomly sampled 20 times from the gene group with the perturbation sensitivity $(S_i) = 1$ (the group is made up of 1526 genes). The 831 genes had a high proportion of annotations like *protein fate, transcription, cell fate, cell type differentiation* and *biogenesis of cellular components*. As shown in Figure 5, these annotations are exclusive to the annotations of the perturbation sensitive genes. And the functional categories are essential processes which are always switched on to keep life go on and it accords with the above mentioned finding that perturbation resistant genes may be 'house keeping' genes.

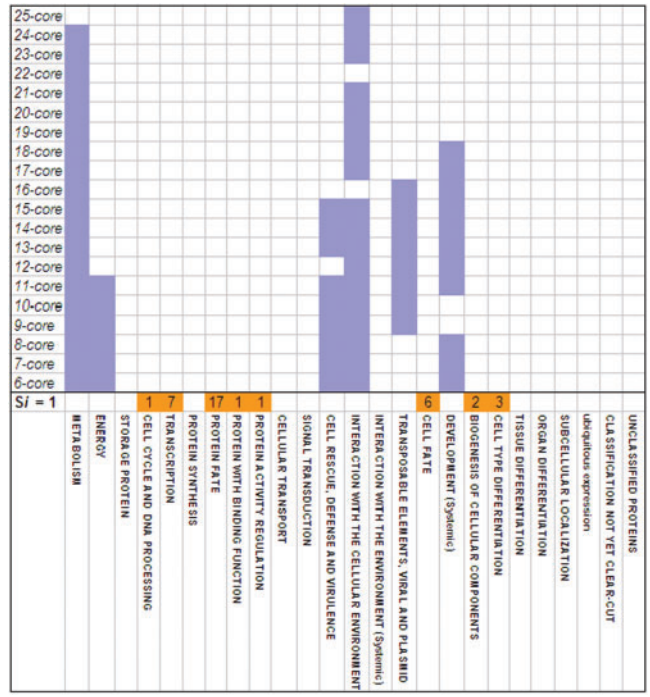

Fig. 4. The distribution of functional categories (P-value < 0.01) in each group of perturbation sensitive (*m-core*) and resistant group ($S_i = 1$) is shown. If a functional category is significantly overrepresented in a group, then the corresponding rectangle is colored. The perturbation sensitive group (blue) and resistant group (orange, the number is the frequency of overrepresentations in 20 random samples of 831 genes) have mutually exclusive functional annotations.

Perturbation Sensitive	⟷	Perturbation Resistant
Non-lethal	⟷	Lethal
Less Physical Interactions	⟷	More Physical Interactions
More Upstream TFs	⟷	Less Upstream TFs
Environmental Interactions & Metabolism	⟷	Essential Cellular Processes

Fig. 5. Summary of correlations among various properties of yeast proteins.

4 DISCUSSION

We conclude from the above exploration that perturbation sensitive genes are usually not essential and have fewer physical interaction partners and their perturbation sensitivity is well correlated with the number of upstream binding transcription factors. Moreover, perturbation sensitive and resistant genes belong to mutually exclusive functional categories. It is summarized in Figure 5.

The 'excess retention' approach (Wuchty and Almaas, 2005) well visualizes the tendency of enrichment or depletion of genes with specific property with the gradual change in perturbation sensitivity gene sets.

The sensitivity in transcriptional changes of a gene following random perturbations has been left relatively under-explored despite being a very interesting and more dynamic property. Through definition of the 'perturbation sensitivity', we add a novel way of characterizing a gene from a genomic perspective to the present genetic annotation system. It has definite biological implications and shows interesting correlations with the present genomic characteristics leading to new biological insights. Several studies have discussed the correlation structures among various genomic characteristics. (Featherstone and Broadie, 2002; Yu *et al.*, 2004) But our study clearly demonstrates them using the 'perturbation sensitivity' as a scaffold for unifying various genomic characteristics.

ACKNOWLEDGEMENTS

This study was supported by a grant from Korea Health 21 R&D Project (A040163) and J.K's educational training was supported by a grant from Korea Health 21 R&D Project (A060711), Ministry of Health and Welfare, Republic of Korea.

Conflict of Interest: none declared.

REFERENCES

Albert,R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.

Featherstone,D.E. and Broadie,K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *BioEssays*, **24**, 267–274.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Markowetz,F. *et al.* (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, **21**, 4026–4032.

Mewes,H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Pe'er,D. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.

Ruepp,A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.

Rung,J. *et al.* (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics*, **18** Suppl. 2), S202–S210.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond.*, **B268**, 1803–1810.

Wasserman,S. and Faust,K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Wuchty,S. and Almaas,E. (2005) Peeling the yeast protein network. *Proteomics*, **5**, 444–449.

Winzeler,E.A. *et al.* (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Yu,H. *et al.* (2004) Genomic analysis of essentiality within protein networks. *Trends Genet*, **20**, 227–231.