PUBLICATION VI

Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast

In: Molecular BioSystems 2009, 5(3):276–287.

Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast†

Peddinti V. Gopalacharyulu,‡[a] Vidya R. Velagapudi,‡[a] Erno Lindfors,[a] Eran Halperin[b] and Matej Orešič*[a]

Received 3rd September 2008, Accepted 5th January 2009
First published as an Advance Article on the web 19th January 2009
DOI: 10.1039/b815347g

In response to environmental challenges, biological systems respond with dynamic adaptive changes in order to maintain the functionality of the system. Such adaptations may lead to cumulative stress over time, possibly leading to global failure of the system. When studying such systems responses, it is therefore important to understand them in system-wide and dynamic context. Here we hypothesize that *dynamic changes* in the topology of functional modules of integrated biological networks reflect their activity under specific environmental challenges. We introduce topological enrichment analysis of functional subnetworks (TEAFS), a method for the analysis of integrated molecular profile and interactome data, which we validated by comprehensive metabolomic analysis of dynamic yeast response under oxidative stress. TEAFS identified activation of multiple stress response related mechanisms, such as lipid metabolism and phospholipid biosynthesis. We identified, among others, a fatty acid elongase IFA38 as a hub protein which was absent at all time points under oxidative stress conditions. The deletion mutant of the IFA38 encoding gene is known for the accumulation of ceramides. By applying a comprehensive metabolomic analysis, we confirmed the increased concentrations over time of ceramides and palmitic acid, a precursor of *de novo* ceramide biosynthesis. Our results imply that the connectivity of the system is being dynamically modulated in response to oxidative stress, progressively leading to the accumulation of (lipo)toxic lipids such as ceramides. Studies of local network topology dynamics can be used to investigate as well as predict the activity of biological processes and the system's responses to environmental challenges and interventions.

Background

Many cellular processes rely upon the concerted action of multiple molecular components. Although such collective phenomena have been difficult to study using traditional methods of molecular biology, analytical methods have been developed over the recent years that afford detection of multiple types of physical and genetic interactions between the components in a high-content and throughput manner.[1] Systems biology provides a conceptual framework to study the complexity of such biological organizations and processes.[2] Using the systems approach, the interactions between the molecular components can be viewed as a complex network whose structure can vary spatially and temporally.[3]

Proteins play a key role in cellular processes and are involved in multiple types of interaction networks. Specific protein–protein interactions play a central role in events such as signal transduction[4] and transcription.[5] An important role of protein–protein interactions has also been recognized in metabolic networks, *e.g.*, in substrate channeling.[6,7] The protein interactome can therefore span multiple and overlapping means of communication, which can occur over a broad range of dynamic and spatial scales.

The global structure of biological networks, *e.g.*, protein–protein interaction[8,9] and metabolic networks,[10] has been extensively studied.[3] Topological analyses of static protein–protein interaction networks in *Saccharomyces cerevisiae* suggest that such networks are scale-free.[11] Therefore, most proteins interact with few partners, while a small but significant fraction of proteins, the ''hubs'', interact with many partners. Hubs in interaction networks are likely to play important roles in regulating biological response. In a study of the allergic immune response of asthma, it was shown that the hubs and superhubs of mouse protein–protein interaction networks exhibit low levels of change in gene expression.[12] This shows the limitation of studying gene expression data in isolation, and shows that concomitant analysis of gene expression data and topological interaction networks provides valuable insights into biological processes.

Not much of the knowledge acquired from the interactome topology studies has been utilized in the analysis and interpretation of high-content molecular profile data. In part, this

[a] *VTT Technical Research Centre of Finland, P.O. Box 1000, Espoo, FI-02044 VTT, Finland. E-mail: matej.oresic@vtt.fi; Fax: +358 20 722 7071; Tel: +358 20 722 4491*
[b] *International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, California, USA. E-mail: heran@icsi.berkeley.edu; Fax: +1 510 666 2956; Tel: +1 510 666 2952*

† Electronic supplementary information (ESI) available: Additional details about methods and results and supplementary metabolomics data with lipidomics raw data, concentrations of lipids and primary metabolites, and relative abundance (%) of fatty acids. See DOI: 10.1039/b815347g

‡ Authors contributed equally.

may be due to the lack of availability of reliable interaction data.[1] Several tools do exist that enable network visualization as well as integration of mRNA expression profiling and interaction network data, including Cytoscape,[13] visANT,[14] genMAPP,[15] PATIKA,[16] and megNet.[17] Pathway analysis methods have been developed such as gene set enrichment analysis (GSEA),[18] aiming to identify functional modules or selected gene sets showing statistically significant and concordant differences across several microarray experiments associated with different biological states. A related method has recently been developed, gene network enrichment analysis (GNEA), which identifies the differences corresponding to subnetworks derived from protein–protein interaction data.[19]

Although GSEA and related methods take into account subtle changes in gene expression that may be missed by traditional tests for differential expression, they rely solely

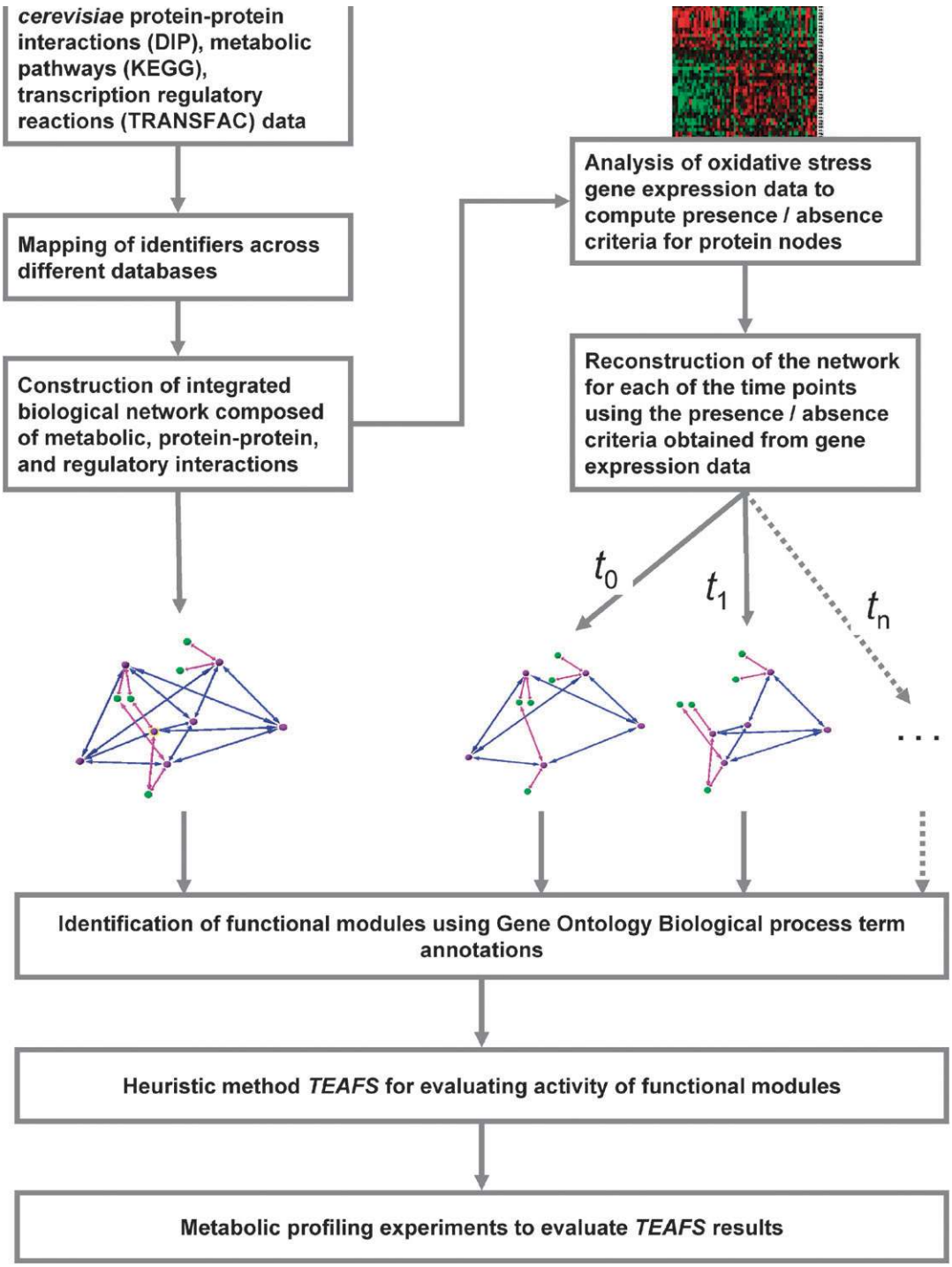

Fig. 1 Outline of the TEAFS method.

on the gene expression levels and ignore the interactions. Specifically, the functional importance of a specific protein within a functional module, and thus its influence on the "output" of the module, is likely to depend on the extent of protein's interactions with other module members. The absence of a protein hub within a functional module may lead to a dysfunctional module, although many of the module proteins may be overexpressed as a compensatory response.

The systems responses to environmental challenges are tightly dynamically regulated, with the aim to maintain the functionality of the system. Change in the expression of a specific gene may lead to a change in the connectivity of the interaction networks by switching some interactions on or off and thus affecting the pathways as well as the products of different functional modules. Such network-level changes are reflected in the changes of topological measures such as average clustering coefficient and average degree.[3] Intuitively, functional modules most relevant for a system's response to environmental changes and interventions may undergo many topological variations. We therefore hypothesize that dynamic changes in topology of integrated biological networks at the functional module level are a sensitive measure of the module activity.

As a way to pursue our hypothesis, we developed a heuristic method for the analysis of integrated molecular profile and interactome data, topological enrichment analysis of functional subnetworks (TEAFS). This method combines time course molecular profile data, such as genome-wide gene expression data, with integrated networks (*i.e.*, protein–protein interaction, metabolic, and regulatory networks). Functional modules in the network are identified using the Gene Ontology[20] and condition specific networks corresponding to all time points are constructed. Changes in module activity are estimated from the changes in topological properties of the module networks (Fig. 1).

Here, we demonstrate the application of TEAFS for the study of dynamic responses to oxidative stress in *S. cerevisiae*.[21] Among all stress conditions, oxidative stress is generally relevant to study because all aerobically growing cells are exposed to reactive oxygen species (ROS) such as the superoxide radical, the hydroxyl radical and hydrogen peroxide, which are all produced during metabolism.[22,23] Moreover, it is well known that oxidative stress is involved in several human pathologies and physiological processes, including cancer, diabetes, cardiovascular disease, ageing and age related diseases. ROS are generated endogenously through leakage of electrons from the mitochondrial respiratory chain to oxygen resulting in superoxide formation.[24] Exposure to ROS can, over time, cause significant damage to macromolecules such as DNA, proteins, carbohydrates and lipids.[25] Cells have evolved different defense mechanisms such as antioxidant enzymes and molecular scavengers to counteract oxidative damage.[26]

As demonstrated in this paper, the application of the TEAFS method to the study of oxidative stress response in yeast proved effective in finding the activity of multiple stress related mechanisms, which the GSEA method could not identify. We validated our findings by performing comprehensive metabolomic analyses of yeast dynamic response under oxidative stress. We conducted the metabolomics experiments by emulating the cell cultivation and hydrogen peroxide treatment protocols identical to the gene expression experiments done by Gasch and colleagues.[21] We profiled lipids using ultra performance liquid chromatography coupled with mass spectrometry (UPLC/MS), primary metabolites using high performance liquid chromatography and mass spectrometry (HPLC/MS) and fatty acids with gas chromatography (GC). To the best of our knowledge, this is the first comprehensive metabolomic analysis of dynamic oxidative stress response in *S. cerevisiae*. Our experimental analyses confirmed the findings of our model and the measured lipid levels correlated with the clustering coefficients of corresponding functional modules.

Results

Network reconstruction

As described in the subsections "Integration of interaction data" and "Integration of gene expression data" in the "Experimental" section, we constructed reference and condition networks corresponding to time points of the oxidative stress gene expression data.[21] The integrated networks contain genes, proteins, metabolites, transcription factors and transcription factor binding sites as the nodes, and protein–protein interactions, metabolic reactions, transcriptional regulatory relationships (*i.e.* transcription factor-binding site interactions, binding site-gene associations), gene encodes protein relationships as the edges. Sizes of these networks are provided in (Table 1).

Hubs

Hubs in a network are the nodes with degrees much larger than the average degree of all nodes in the network. We selected top 15 high degree nodes as hubs. We identified the protein node JSN1/PUF1 (Uniprot: P47135), which belongs to the Puf family of mRNA binding proteins,[27] as a hub in most of the networks, except at time points 40 and 100 min. The microsomal fatty acid elongase 3-ketoreductase (IFA38; UniProt ID: P38286), encoded by YBR159w gene, was a hub protein that was absent at all time points (Table 2).

Most hub proteins are related to environmental stress responses and DNA repair mechanisms (Table 2). Selected examples of some other hubs are P29295 (HRR25)—casein kinase I, P39743 (RVS167)—reduced viability upon starvation protein and P06787 (CMD1)—Calmodulin (see ESI†). Most of the reference network hub proteins were absent in 40 and 100 min networks. Interestingly, at these two time points, most of the identified hubs were subunits of DNA directed RNA polymerases I, II and III, which are involved in transcription and ribosome biogenesis.

Connectivity of biological process modules is regulated by oxidative stress

We identified a total of 659 biological process modules in the reference and condition networks. Clustering coefficients were calculated for each module in each condition and a *t*-test was performed to select modules with average clustering

Table 1 The number of nodes and edges of the reference and condition networks

Time/min	Proteins	Protein–protein interactions	TF–TF interactions	TF–BS binding relations	Gene-encodes-protein relations	Biochemical reactions
10	2412	3612	7	329	5080	881
20	2555	3611	11	269	5445	880
30	3799	7873	15	436	7823	1306
40	2784	4451	9	249	5797	997
50	3115	5913	14	325	6566	1195
60	3372	7265	15	397	7147	1338
80	3943	9504	22	488	8267	1552
100	3165	7265	15	397	7147	1338
120	3366	6656	20	353	6749	1380
160	4138	10555	29	430	8601	1618
Reference	5695	17323	52	666	11663	1936

Abbreviations: TF, Transcription Factor; BS, Binding Site.

coefficients significantly greater than zero. A total of 346 modules whose average clustering coefficient over time was significantly greater than zero were selected for further study of topological changes. The clustering coefficients of the protective mechanisms against oxidative stress, in particular base excision repair, defense mechanisms like DNA repair, mismatch repair, regulation of meiosis, peroxisome matrix protein import have shown varying profiles. Nucleic acid metabolism and regulation of DNA replication modules also showed varying profiles. Modules that are related to double strand break repair mechanisms *via* non-homologous end joining, DNA repair and DNA synthesis during repair varied with respect to the reference network (Fig. 2). A fluctuating clustering coefficient was also observed for modules related to other stress responses, such as osmotic stress and salt stress, endoplasmic reticulum (ER) associated protein catabolism, response to cadmium ion, and osmosensory signaling pathway.

The overall connectivity of lipid metabolism and phospholipid biosynthesis modules decreased in time. Phospholipid biosynthesis module showed a drop in connectivity at 40 min (Fig. 2). The absence of most of the hubs in the integrated interaction network at these time points (Table 2) has perhaps led to this drop in connectivity.

Visualization of the module networks is an alternative way to understand the dynamic changes of the modules. For a selected module from each category in Fig. 2, we present the network view in Supplementary Fig. S2.†

Topological enrichment of functional subnetworks

We identified differentially active modules using a false discovery rate (FDR) q-value cut-off of 0.05. FDR controls the expected proportion of type I errors. A total number of 174, 162, and 125 differentially active modules were identified based on enrichment scores EDA(LIC) (*i.e.*, extent of differential activity for local in-connectivity), EDA(LOC) (*i.e.*, extent of differential activity for local out-connectivity), and EDA(LC) (*i.e.*, extent of differential activity for local clustering), respectively. These account for approximately 220 distinct modules among the 346 modules which were analyzed. Most of the differentially active modules are involved in the environmental stress responses, oxidative stress response in particular (Supplementary Table S2†). The most active modules identified by TEAFS include regulation of cell cycle and check points, response to DNA damage stimulus (*i.e.*, repair mechanisms), cell wall organization, pentose phosphate shunt, biosynthesis of stress protectors (*i.e.*, glycogen and trehalose), signal transduction pathways, post-translational modifications, regulation of transcription and vacuolar acidification (Table 3).

Comparison of TEAFS with GSEA

In order to compare our integrative analysis method with existing pathway analysis methods, we chose the commonly utilized tool called gene set enrichment analysis (GSEA).[18] GSEA revealed that none of the gene sets scored a positive enrichment score (up-regulated) and one gene set, GO: 0030437 (sporulation; *sensu* Fungi), scored a negative enrichment score (down-regulated) at the recommended FDR q-value < 0.25 threshold.

With the nominal p-value cut-off (<0.05), GSEA detected 12 up-regulated and 30 down-regulated gene sets. Down-regulated gene sets include (1) lipid metabolism (2) nucleus export mechanisms, namely mRNA, rRNA, snRNA, tRNA, and protein–nucleus export (3) nucleus import mechanisms, namely NLS-bearing substrate, snRNP protein, ribosomal protein, and protein–nucleus import (4) ribosome biogenesis (5) nuclear pore organization and biogenesis (6) DNA-dependent regulation transcription and (7) removal of non-homologous ends, among others. Up-regulated gene sets include (1) vesicle fusion and (2) SRP-dependent co-translational protein-membrane targeting. All these processes were identified to be significantly active modules in TEAFS analysis as well. Results of GSEA are included in the Supplementary Table S2.†

Dynamic metabolite profiles of oxidative stress response in yeast

In order to validate our findings from TEAFS analysis, we performed metabolomic analysis of samples obtained from a time course study of *S. cerevisiae* under oxidative stress. By applying a global lipidomic analysis, we identified and quantified a total of 263 lipid molecular species representing multiple lipid classes. The main intermediary metabolites of central carbon metabolism (*i.e.*, glycolysis, TCA cycle and pentose

Table 2 Hubs in the reference network. The time points where these hubs become absent (removed from the network) are provided. The proteins are ordered according to the number of interacting partners in the reference network. Descriptions are from *Saccharomyces* Genome Database[57]

Uniprot ID	Name	Description	Absent at time points/minutes
P47135	Protein JSN1; Protein PUF1	Member of the Puf family of RNA-binding proteins, interacts with mRNAs encoding membrane-associated proteins; overexpression suppresses a tub2-150 mutation and causes increased sensitivity to benomyl in wild-type cells	40, 100
Q02821	Importin alpha subunit; Karyopherin alpha subunit; Serine-rich RNA polymerase I suppressor protein	Karyopherin alpha homolog, forms a dimer with karyopherin beta Kap95p to mediate import of nuclear proteins, binds the nuclear localization signal of the substrate during import; may also play a role in regulation of protein degradation	20
P00546	Cell division control protein 28	Catalytic subunit of the main cell cycle cyclin-dependent kinase (CDK); alternately associates with G1 cyclins (CLNs) and G2/M cyclins (CLBs) which direct the CDK to specific substrates	20, 40, 50, 60
Q12349	ATP synthase H chain, mitochondrial precursor	Subunit h of the F0 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis	20, 30
P07703	DNA-directed RNA polymerases I and III 40 kDa polypeptide	RNA polymerase subunit, common to RNA polymerase I and III	120
P38987	Protein TEM1	GTP-binding protein of the ras superfamily involved in termination of M-phase; controls actomyosin and septin dynamics during cytokinesis	10, 20, 30, 40, 50, 60, 120
P32366	Vacuolar ATP synthase subunit d; V-ATPase d subunit; Vacuolar proton pump d subunit; V-ATPase 39 kDa subunit; V-ATPase subunit M39	Subunit d of the five-subunit V0 integral membrane domain of vacuolar H + -ATPase (V-ATPase), an electrogenic proton pump found in the endomembrane system; stabilizes VO subunits; required for V1 domain assembly on the vacuolar membrane	40
Q02630	Nucleoporin NUP116/ NSP116; Nuclear pore protein NUP116/NSP116	Subunit of the nuclear pore complex (NPC) that is localized to both sides of the pore; contains a repetitive GLFG motif that interacts with mRNA export factor Mex67p and with karyopherin Kap95p; homologous to Nup100p	10, 20, 30, 40, 50, 60, 80
P32569	RNA polymerase II mediator complex subunit 17; Suppressor of RNA polymerase B 4	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation	40
P38264	Inorganic phosphate transporter PHO88	Probable membrane protein, involved in phosphate transport; pho88 pho86 double null mutant exhibits enhanced synthesis of repressible acid phosphatase at high inorganic phosphate concentrations	20, 30
P29055	Transcription initiation factor IIB; General transcription factor TFIIB; Transcription factor E	Transcription factor TFIIB, a general transcription factor required for transcription initiation and start site selection by RNA polymerase II	10, 20, 30, 40
P38822	Protein BZZ1	SH3 domain protein implicated in the regulation of actin polymerization, able to recruit actin polymerization machinery through its SH3 domains, colocalizes with cortical actin patches and Las17p, interacts with type I myosins	10, 20, 40, 50, 60, 100
P38286	Putative oxidoreductase YBR159W	Microsomal beta-keto-reductase; contains oleate response element (ORE) sequence in the promoter region; mutants exhibit reduced VLCFA synthesis, accumulate high levels of dihydrosphingosine, phytosphingosine and medium-chain ceramides	10, 20, 30, 40, 50, 60, 80, 100, 120, 160
P27999	DNA-directed RNA polymerase II subunit 9; DNA-directed RNA polymerase II 14.2 kDa polypeptide	RNA polymerase II subunit B12.6; contacts DNA; mutations affect transcription start site; involved in telomere maintenance	10, 20, 30, 50, 60, 80, 100, 120
P02309	Histone H4	One of two identical histone H4 proteins (see also HHF2); core histone required for chromatin assembly and chromosome function; contributes to telomeric silencing; *N*-terminal domain involved in maintaining genomic integrity	
P40054	D-3-phosphoglycerate dehydrogenase 1; 3-PGDH 1	3-phosphoglycerate dehydrogenase, catalyzes the first step in serine and glycine biosynthesis; isozyme of Ser33p	10, 20, 40, 120

phosphate pathway) and fatty acids were also measured. Consistent increases in the levels of trehalose-6-phosphate ($p = 0.00056$) and decreases in that of pyruvate ($p = 0.000052$) and mannose-6-phosphate ($p = 0.048$) were observed during oxidative stress with respect to the *S. cerevisiae* under normal conditions. The major change observed in fatty acid profiles was the consistent increase in the levels of palmitic acid (C16:0) during oxidative stress ($p = 0.00027$) with respect to *S. crevisiae* under normal conditions. Moreover, the relative palmitate concentration increased over time (Pearson correlation $r = 0.75$, $p = 0.008$) (Fig. 3B).

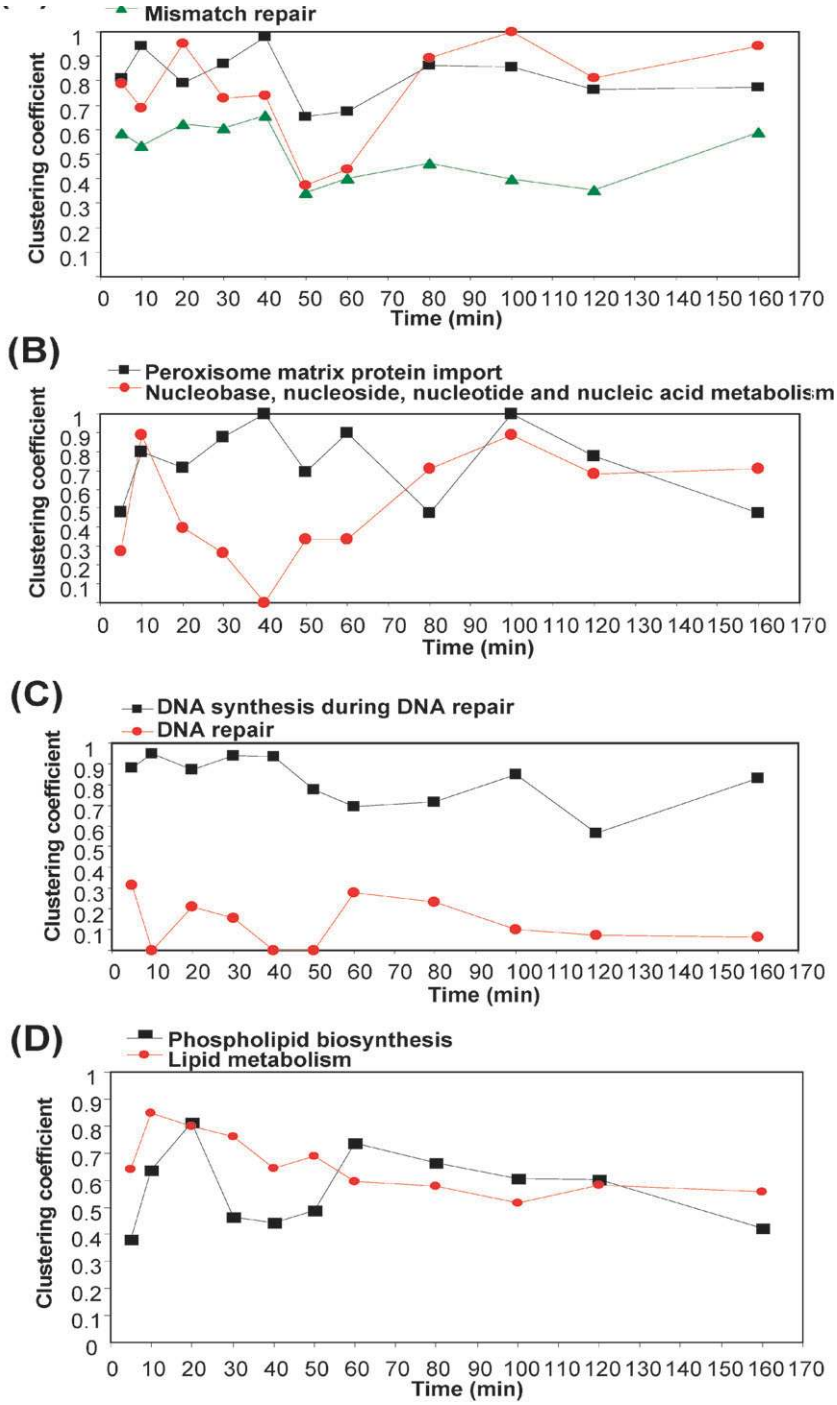

Fig. 2 Changes in the connectivity of selected biological process modules: (A) Nucleotide repair mechanisms, (B) DNA metabolism, (C) double strand break repair mechanisms, and (D) lipid metabolism of the reference and condition networks over time period.

We compared the lipid metabolism related changes from topological and metabolomic analyses. An increase in average ceramide concentrations was observed over the time period ($r = 0.8$, $p = 0.003$) (Fig. 3A). The correlation between the average phospholipid concentrations and phospholipid biosynthesis module was positive ($r = 0.604$) and marginally significant ($p = 0.064$) (Fig. 4).

Data from metabolomics experiments are provided as ESI.†

Discussion

Dynamic topology changes of functional subnetworks in oxidative stress response

The aim of this study was to explore the hypothesis that the dynamic changes in the topology of integrated biological networks at the functional module level are a reasonable measure for the module activity. Motivated by this hypothesis,

Table 3 Selected results of heuristic method TEAFS, which identifies module activity based on change in the local topological properties

	Local in connectivity (LIC)		Local out connectivity (LOC)		Local clustering (LC)		GSEA results	
Module name	EDA	FDR	EDA	FDR	EDA	FDR	ES	*P*
G1/S transition of mitotic cell cycle	3176.25	0.01536	3179.05	0.02638	15.25	0	NA	NA
Response to DNA damage stimulus	1738.66	0	1738.66	0	1.25	0.00097	NA	NA
Cytokinesis	1573.19	0.03329	1489.44	0.05171	47.86	0	NA	NA
Cell wall organization and biogenesis	1562.55	0	1338.46	0	24.06	0.02369	NA	NA
Response to osmotic stress	1397.20	0	1364.24	0	11	1	NA	NA
Vacuolar acidification	1053.22	0	1486.41	0.00083	28.98	0.00882	NA	NA
Glycogen metabolism	776.54	0.0004	777.51	0.00022	4.86	1	NA	NA
Signal transduction	739.31	0	763.33	0	12.54	0	NA	NA
Cellular morphogenesis during vegetative growth	648.92	0.01586	648.92	0.01428	NA	NA	NA	NA
Invasive growth (*sensu Saccharomyces*)	405.96	0.00078	367.06	0.23532	3.3	0.14813	NA	NA
Response to salt stress	348.11	0	349.44	0	NA	NA	NA	NA
Methionine metabolism	264.47	0.00021	240.05	0.0024	8.07	0.02253	NA	NA
Regulation of cell cycle	243.35	0.0068	245.39	0.01437	5.21	0.55413	NA	NA
Ceramide biosynthesis	183.39	0.05922	183.39	0.064	NA	NA	NA	NA
Pentose-phosphate shunt	137.18	0	237.59	0	15.69	0	NA	NA
Regulation of transcription	123.22	0.01096	123.47	0.02475	NA	NA	NA	NA
N-terminal protein myristoylation	62.93	0.03329	62.93	0.03557	NA	NA	NA	NA
Thiamin biosynthesis	55.98	0.0004	55.98	0.00022	NA	NA	NA	NA
Lipid metabolism	54.45	0.04958	54.45	0.05263	NA	NA	−0.554	0.04131
Protein sumoylation	47.03	0.05922	47.03	0.06354	NA	NA	NA	NA
Sulfur metabolism	30.83	0	27.59	0	0.15	0	NA	NA

Abbreviations: EDA, Extent of Differential Activity; FDR, False Discovery Rate *q* value; GSEA, Gene Set Enrichment Analysis; ES, Enrichment Score; *P*, Nominal *p* value; NA, not significant.

we developed a method for the analysis of integrated interactome and molecular profile data, topological enrichment analysis of functional subnetworks (TEAFS). Our combined findings from the TEAFS analysis of gene expression data from yeast oxidative stress response time series experiments[21] and from comprehensive metabolomic analysis tend to support our hypothesis, as well as suggest that the application of topological measures, such as the clustering coefficient, is a useful way of estimating module activity. A highly active module found by TEAFS should be understood as a most changing module rather than as an up-regulated module during oxidative stress.

A number of active modules identified by TEAFS analysis have been previously associated with environmental stress responses in general, and in particular to the oxidative stress. The topmost active module is "G1/S transition of mitotic cell cycle", although the TEAFS method also identified several other modules related to the cell cycle. It is well known that oxidative stress causes damage to the cell by delaying the transition from G1 to S phase and subsequently arresting the cell cycle.[28] This delay in the progression of the cell cycle might allow time for the repair processes to take place.

Pentose phosphate shunt

Since oxidative stress causes a shift in the redox state [NAD(P)H/NAD(P)] of the cell, one would expect a change in the activity of a module that has a role in maintaining the redox homeostasis of the cell. Indeed, our results showed "pentose phosphate shunt" as an active module. The pentose phosphate pathway is the main source of cellular reducing power in the form of NADPH and plays critical role in maintaining the redox balance of the cell. Reducing equivalents in the form of NADPH are not only essential for biosynthesis of cellular macromolecules but also required for the action of many antioxidant enzymes.[29] The shift towards the pentose phosphate pathway under oxidative stress may also be needed to provide ribonucleotides in order to repair DNA and increase proliferation.[30]

In response to diverse stress stimuli, the disaccharide trehalose is produced in large quantities to protect the cell against denaturation and aggregation of proteins[31] and also enhances the buffer capacity of the cell to manage osmotic instability and energy reserves.[21] Consistent with previous findings, our primary metabolite analysis revealed significant increases in the levels of trehalose-6-phosphate, a precursor of trehalose biosynthesis, in stress conditions *vs.* controls.[21] A continuous decrease in trehalose-6-phosphate concentration was observed over time during oxidative stress. Such a trend is likely due to the co-induction of both biosynthetic (trehalose-6-phosphate synthase) and catabolic (trehalase) enzymes of trehalose in response to environmental stimuli.[21] This supports our finding that transcriptional activation of the trehalose biosynthetic gene in response to stress stimulus was not accompanied by a corresponding accumulation of trehalose-6-phosphate, suggesting futile recycling of trehalose resulted instead of accumulation.[32]

Lipid metabolism

Lipids are known to play important roles in oxidative stress.[33] TEAFS detected changes in lipid related modules such as lipid

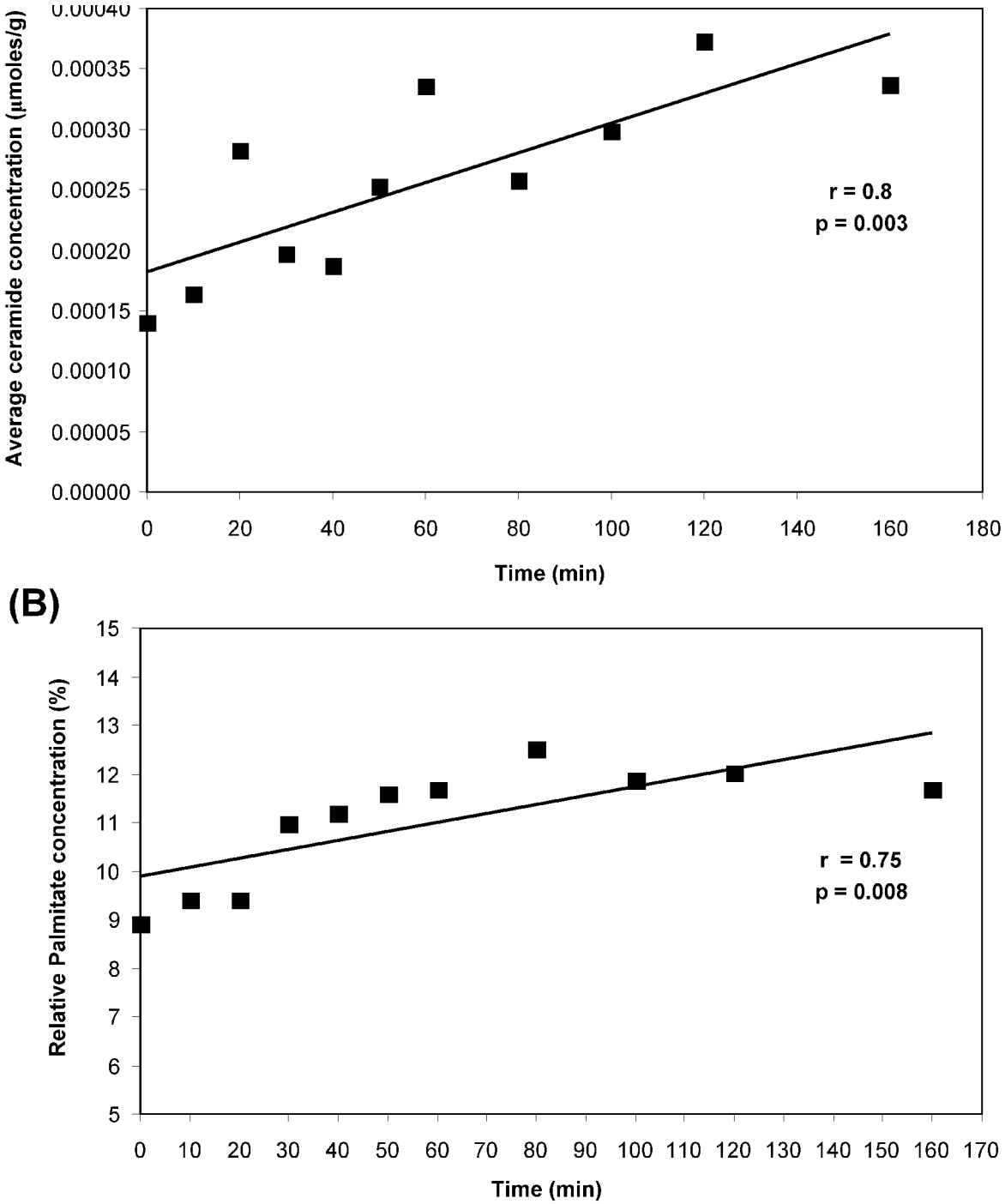

Fig. 3 Increasing trend of average ceramide concentration and relative palmitate concentration.

metabolism and phospholipid biosynthesis during oxidative stress. Changes in the ceramide biosynthesis module were also marginally significant (FDR $q = 0.059$). Consistent with the results of TEAFS analysis, lipidomic analysis revealed significant changes in the lipidome, particularly in ceramide and phospholipid levels, during oxidative stress.

Plasma membrane phospholipids are vulnerable sites and major targets of oxidative stress, which causes damage to the cell by changing the structural organisation and membrane permeability. The cytotoxic effect of hydrogen peroxide causes disruption of phospholipids by generating lipoperoxides by peroxidation of the fatty acid component.[34] As the average phospholipid levels are reduced in oxidative stress, our results are in accordance with these findings.

The observed increase in ceramide concentrations over time is in accordance with the previous findings that the cell responds to diverse stresses with ceramide generation.[35] Ceramides, a class of sphingolipids, are secondary messengers

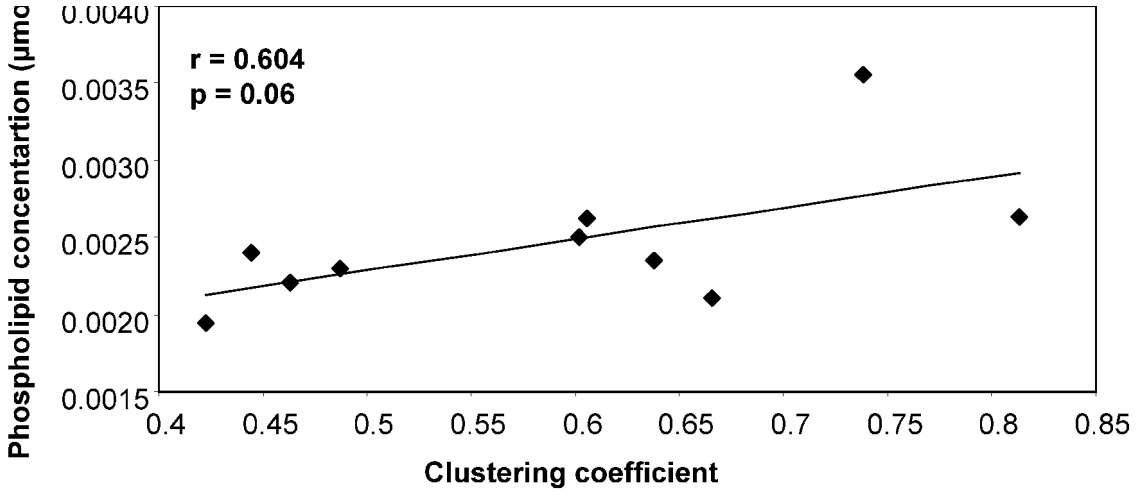

Fig. 4 Clustering coefficients of the phospholipid biosynthesis module and phospholipid concentrations.

for various cellular functions and act as signalling molecules in oxidative stress.[36,37] Ceramides can be synthesized either by hydrolysis of sphingomyelins by sphingomyelinases (SMase) or *de novo* synthesis by ceramide synthase.[38] Some of the enzymes involved in ceramide generation have been shown to be induced by environmental stress stimuli and signal from damaged DNA, for example double strand breaks, which is the most common consequence of oxidative stress.[39]

Hub proteins affected by oxidative stress

Our fatty acid analysis results showed a significant increase in the levels of palmitic acid (16 : 0), a precursor of *de novo* ceramide biosynthesis.[40] As one of the major findings from topological analysis of gene expression data, the fatty acid elongase 3-ketoreductase (IFA38) which is encoded by the YBR159w gene[41] and was identified as one of the hub proteins in our integrated networks, was absent at all time points under oxidative stress (Table 2). The YBR159w mutant exhibits a phenotype common to other mutants with defects in fatty acid elongation, including accumulation of ceramides and related reactive sphingolipids.[41] It is therefore an intriguing possibility that accumulation of palmitate, a substrate to elongase system involving IFA38, and subsequent accumulation of ceramides are in part consequences of IFA38 response to oxidative stress.

Interestingly, the human homolog of IFA38, type 12 17β-hydroxysteroid dehydrogenase (HSD17B12; Uniprot ID: Q53GQ0), is also involved in fatty acid elongation.[42] Although poorly characterized so far, recently it was shown that HSD17B12 is also a major estrogenic 17β-hydroxysteroid dehydrogenase responsible for the conversion of estrone into estradiol in women[43] and that it is highly expressed in the human breast cancer tissue.[44] HSD17B12 clearly needs to be investigated further in the context of oxidative stress as well as crosstalk of lipid and estrogen metabolism.

A hub protein, which was present at most of the time points, was JSN1 from the Puf family of mRNA binding proteins.[27] Puf proteins in *S. cerevisiae* belong to a structurally related family of cytoplasmic RNA binding proteins that are implicated in developmental processes in various eukaryotes.[45] The human homolog of JSN1, Pumilio-2 (PUM2), is expressed predominantly in human embryonic stem cells and germ cells, and is required for germ cell development.[46] Although both JSN1 and PUM2 are poorly characterized, it appears that both in yeast and in human these two homologous proteins act as translational regulators, interacting with multiple other proteins.

Comparison of TEAFS and GSEA

A large number of significantly active modules found by TEAFS relevant to the oxidative stress response were not found in the GSEA analysis. Note that these two methods have inherent differences. TEAFS takes interaction network structure into account by measuring the variation of the local connectivity and clustering, whereas GSEA is based on membership of molecules in gene sets. The TEAFS method uses gene expression data qualitatively, because the presence/absence criteria gives raise to exactly two classes of proteins nodes: present or absent, whereas GSEA takes the sign and magnitude of the differential expression into account.

Notably, GSEA identified only one gene set to be affected by the oxidative stress using the recommended FDR significance cut-off. With the nominal p value cut-off it identified only a small number of affected modules, which may be unlikely because oxidative stress affects many biological processes at transcriptional level. In contrast, TEAFS identified multiple known oxidative stress-related responses that could be verified based on existing literature. Moreover, within our capabilities we confirmed some changes suggested by TEAFS using comprehensive metabolomic profiling experiments. Taken together, these observations show that TEAFS is more sensitive in detecting active pathways than GSEA. The ideal way to assess the accuracy of TEAF method would involve performing extensive literature survey or a large number of experiments to confirm the changes suggested by TEAFS. This, however, is infeasible with our limited resources.

The ability of TEAFS to detect multiple changes related to oxidative stress response thus suggests that change in module

connectivity is a better measure of its activity than changes in the levels of gene expression. It is possible, with some modifications, to apply our methodology to analyze gene expression data from different experimental designs. Presence/absence criteria that we employed here for reconstructing modules are specific to cDNA platforms and need to be properly adopted for use with other platforms. Moreover, if available, it would be advantageous to use protein expression information rather than mRNA expressions. Likewise, instead of GO, other methods for module definition could be applied, *e.g.*, by using other biological knowledge such as KEGG metabolic pathways or alternatively by applying computational module identification methods.[47,48]

Experimental

Integration of interaction data

We previously presented an approach for heterogeneous data integration.[17,49] Our software presented there, called megNet, is capable of integrating data from multiple types of biological databases from an in-house database system and represent the results as integrated networks. Here, we used megNet to construct an *S. cerevisiae* genome-wide integrated interaction network by obtaining information from multiple interaction sources. We used *S. cerevisiae* protein–protein interaction data from the January 2006 release of DIP,[50] 'gene encodes protein' relationships from the EMBL nucleotide sequence database,[51] metabolic interactions from the 9 February 2006 release of KEGG[52] and transcriptional regulatory relationships from release 9.4 of Transfac.[53]

Integration of gene expression data

In this study, we used a genome-wide gene expression time course data of the oxidative stress response from a study of responses of *S. cerevisiae* cells to various environmental changes.[21] The oxidative stress data is based on dual channel cDNA microarray experiments and provides gene expression measurements at 10, 20, 30, 40, 50, 60, 80, 100, 120, and 160 min after exposing yeast to oxidative stress. In order to integrate the gene expression information into the integrated network, we first constructed a reference network by removing from it the protein nodes (and their links) whose transcripts were not analyzed in the oxidative stress experiment. Next, we used the reference network as the starting point to reconstruct the condition specific networks (time point networks) corresponding to each time point of the oxidative stress experiment by applying presence/absence criteria as follows.

In order to construct a condition specific network, we defined presence/absence criteria for proteins based on the expression of genes encoding them as described by Luscombe and colleages.[54] Briefly, we first classified the first channel intensity (*i.e.*, reference condition mRNA intensity) values as high, medium and low by using *k*-means clustering of the log transformed values. Each protein was designated as present or absent based on this first channel intensity level and whether the gene was up, constant or down-regulated in the stress condition, as indicated by the corresponding expression log ratio being positive, zero or negative, respectively (Table 4).

Table 4 Criteria for the presence and absence of proteins. The reference intensity levels are divided into high, medium or low regions based on *k*-means clustering. A gene is up, down or constant based on its log ratio being positive, negative, or zero. During the reconstruction of the network, proteins encoded by the absent genes are removed from the network

Reference condition intensity value level	Change in expression level during stress condition	Conclusion on presence or absence
High	Up	Present
	Constant	Present
	Down	Present
Medium	Up	Present
	Constant	Present
	Down	Absent
Low	Up	Present
	Constant	Absent
	Down	Absent

We mapped each spot on the array to the corresponding protein by using the systematic name of the gene, also known as ordered locus name (OLN). Each OLN was translated to its corresponding Swissprot protein accession number using the index of *S. cerevisiae* entries and their corresponding gene designations.[55] A condition specific (*i.e.* time point) network is obtained by removing the absent proteins and their links from the reference network using this criterion (Table 4).

Identification of biological process modules

We used Gene Ontology (GO)[20] biological processes for the identification of the biological process modules. UniProt protein database[55] provides GO term annotations for each protein. These annotations were obtained by querying our in-house databases. A biological process module in an integrated network is identified as the interaction neighbourhood of the proteins annotated by that GO biological process term. We constructed the modules from the reference and time point networks by identifying the network neighbourhood in the corresponding integrated network.

Topological enrichment analysis of functional subnetworks (TEAFS)

TEAFS consists of two main steps. In the first step, we compute enrichment scores which represent the overall amount of topological variation within a module over the course of oxidative stress time points. In the second step, we perform a permutation test to assess the statistical significance of the enrichment score.

For each functional module, the extent of differential activity (EDA) enrichment score is calculated for three topological parameters: local in- or out-connectivity (LIC, LOC respectively) and local clustering (LC).

Local in- and out-connectivity, LIC(m, t) and LOC(m, t), of a module m at time point t is the sum of in- and out-degrees of proteins that are present at time point t, respectively. Local clustering LC(m, t) is the sum of clustering coefficients of the proteins present at time point t. LIC(m, 0), LOC(m, 0) and LC(m, 0) represent the sum of degrees or clustering coefficients of all proteins in the module m, because the time point 0 represents the functional module in the reference network.

Extent of differential activity EDA(m) is the standard deviation over time t of LIC(m, t), LOC(m, t) and LC(m, t). For simplicity we call each of these EDAs as EDA(LIC, m), EDA(LOC, m) and EDA(LC, m), respectively.

In summary, EDA corresponding to LIC, LOC, and LC are defined as

$$\mathrm{EDA(LIC},\, m) \,=\, \sigma_t(\mathrm{LIC}(m,t)) \qquad (1)$$

$$\mathrm{EDA(LOC},\, m) \,=\, \sigma_t(\mathrm{LOC}(m,t)) \qquad (2)$$

and

$$\mathrm{EDA(LC},\, m) \,=\, \sigma_t(\mathrm{LC}(m,t)) \qquad (3)$$

In order to assess the statistical significance of the enrichment score we calculated p values for each module by a permutation test. In each permutation, we randomly deactivated proteins in the module network based on deactivation ratios. The deactivation ratio $x(t)$ at time point t is defined as the ratio between sum of a topological measure of the proteins that are present at time t while absent at $t + 1$, and sum of the topological measure of the proteins that are present at time t. Here, the topological measure stands for in-degree, out-degree or clustering coefficient, depending on the case. Our null model therefore assumes that the proteins are being deactivated uniformly across all networks. We generated 10 000 such permutations. The p-value for a module's EDA is the fraction of the permutations which have at least as much EDA as that of the real module network. In order to account for the multiple comparisons, we applied Bonferroni correction to the p-values as well as computed false discovery rate (FDR) q-values.

High EDA values represent high variation in the topological parameters (LIC, LOC or LC). High variation in topological parameters means that the module network underwent larger topological changes, and hence is very active under the oxidative stress condition. Modules with high EDA values and corresponding FDR q-value < 0.05 are called significantly topologically enriched modules or simply highly active modules.

Yeast experiment

Yeast strain CEN.PK was used in this study. Complete experimental details are described in Supplementary Protocol S1.† In brief, we performed the main cultivation and induced oxidative stress *via* hydrogen peroxide treatment by following the protocol identical to the one described by Gasch and colleagues.[21] Cells were grown to the mid-log phase and 20 ml of culture was taken out at time zero as a reference and 20 ml of pre warmed YPD medium supplemented with H_2O_2 was added to give a final concentration of 0.3 mM H_2O_2. Samples were collected at 10, 20, 30, 40, 50, 60, 80, 100, 120 and 160 min. The same procedure was followed for controls (*i.e.*, without H_2O_2). We could not afford any biological or technical replicates, because of the quick quenching for primary metabolite analysis, and collection of samples for fatty acid and lipid analyses was within a time interval of 10 minutes, simultaneously for both cases and the controls. Lipids were extracted using a chloroform–methanol solvent and analysed with ultra performce liquid chromatography (UPLC™) coupled to a Q-ToF Premier mass spectrometer (Waters, Inc., Milford, MA). The raw data was converted into netCDF files using Dbridge software from MassLynx (Waters, Inc., Milford, MA) and processed using MZmine software version 0.60.[56] In order to analyse primary metabolites, samples were quickly quenched in cold methanol and metabolites were extracted using boiling ethanol protocol. The glycolytic phosphorous and TCA-cycle compounds were detected using high performance liquid chromatography (HPLC) coupled to a Quattro Micro triple quadrupole mass spectrometer (Waters, Inc., Milford, MA) combined with HPLC-MS. Fatty acids were extracted using absolute methanol solvent and GC was carried out using a HP 5890 GC-FID System equipped with HP-FFAP that was directly connected to a FID detector (Agilent Technologies, Waldbronn, Germany). Additional experimental details are provided in Protocol S1.†

Gene set enrichment analysis (GSEA)

We defined gene sets using GO biological process term annotations. We developed a program in the Java programming language to query the GO database and to generate a gene set database in gene matrix transposed (GMT) file format. This program accessed the GO database from our bioinformatics system. We formatted the stress data in gene cluster text (GCT) file format and created the phenotype data in continuous file format (CLS) by entering numerical values for time points in hours. The analysis was performed using the standalone GSEA Java software. Description of all the file formats and the GSEA software are available freely at the program website (http://www.broad.mit.edu/gsea/). Parameters used for the analysis are provided in the Supplementary Table S3.†

Conclusions

Our results imply that the connectivity of the system is being dynamically modulated in response to oxidative stress, leading to progressive accumulation of (lipo)toxic lipids such as ceramides. Our approach differs from conventional strategies for pathway analysis in two ways: (1) by achieving integration of heterogeneous data and (2) by taking functional module connectivity into account. The findings from the study of the oxidative stress response in yeast presented in this paper stand as a proof-of-concept for the applicability of our network analysis strategy as well as supporting the hypothesis that dynamic connectivity changes of the system in response to stress or other external stimuli may determine biological function. The study of local network topology dynamics can therefore be used as an effective tool to investigate as well as predict the activity of biological processes and system's responses to environmental challenges and interventions.

Acknowledgements

We thank John Londesborough, Antonio Vidal-Puig and Sampsa Hautaniemi for valuable comments and discussions, Tuulikki Seppänen-Laakso, Anna-Liisa Ruskeepää, Helena Simolin and Eija Rintala for technical assistance, and Laxman Yetukuri for help with lipid identification. This work is a part of the research program "White Biotechnology—Green

Chemistry" (Academy of Finland; Finnish Centre of Excellence programme, 2008-2013, Decision number 118573).

References

1 A. Beyer, S. Bandyopadhyay and T. Ideker, *Nat. Rev. Genet.*, 2007, **8**, 699–710.
2 T. Ideker, T. Galitski and L. Hood, *Annu. Rev. Genomics Hum. Genet.*, 2001, **2**, 343–372.
3 A.-L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
4 G. B. Cohen, R. Ren and D. Baltimore, *Cell*, 1995, **80**, 237–248.
5 L. O. Barrera and B. Ren, *Curr. Opin. Cell Biol.*, 2006, **18**, 291–298.
6 E. W. Miles, S. Rhee and D. R. Davies, *J. Biol. Chem.*, 1999, **274**, 12193–12196.
7 S. Y. Tsuji, N. Wu and C. Khosla, *Biochemistry*, 2001, **40**, 2317–2325.
8 H. Jeong, S. P. Mason, A.-L. Barabasi and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
9 S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill and M. Vidal, *Science*, 2004, **303**, 540–543.
10 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabasi, *Nature*, 2000, **407**, 651–654.
11 S.-H. Yook, Z. N. Oltvai and A.-L. Barabási, *Proteomics*, 2004, **4**, 928–942.
12 X. Lu, V. V. Jain, P. W. Finn and D. L. Perkins, *Mol. Syst. Biol.*, 2007, **3**, e98.
13 M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader, *Nat. Protoc.*, 2007, **2**, 2366–2382.
14 Z. Hu, D. M. Ng, T. Yamada, C. Chen, S. Kawashima, J. Mellor, B. Linghu, M. Kanehisa, J. M. Stuart and C. DeLisi, *Nucleic Acids Res.*, 2007, **35**, W625–632.
15 N. Salomonis, K. Hanspers, A. Zambon, K. Vranizan, S. Lawlor, K. Dahlquist, S. Doniger, J. Stuart, B. Conklin and A. Pico, *BMC Bioinformatics*, 2007, **8**, 217.
16 U. Dogrusoz, E. Z. Erson, E. Giral, E. Demir, O. Babur, A. Cetintas and R. Colak, *Bioinformatics*, 2006, **22**, 374–375.
17 P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmen and M. Oresic, *Bioinformatics*, 2005, **21**, i177–185.
18 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.
19 M. Liu, A. Liberzon, S. W. Kong, W. R. Lai, P. J. Park, I. S. Kohane and S. Kasif, *PLoS Genet.*, 2007, **3**, e96.
20 M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight and J. Eppig, *Nat. Genet.*, 2000, **25**, 25–29.
21 A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown, *Mol. Biol. Cell*, 2000, **11**, 4241–4257.
22 D. J. Jamieson, *Yeast*, 1998, **14**, 1511–1527.
23 K. P. Mahon, T. B. Potocky, D. Blair, M. D. Roy, K. M. Stewart, T. C. Chiles and S. O. Kelley, *Chem. Biol.*, 2007, **14**, 923–930.
24 M. S. Cooke, M. D. Evans, M. Dizdaroglu and J. Lunec, *FASEB J.*, 2003, **17**, 1195–1214.
25 T. B. Salmon, B. A. Evert, B. Song and P. W. Doetsch, *Nucleic Acids Res.*, 2004, **32**, 3712–3723.
26 G. W. Thorpe, C. S. Fong, N. Alic, V. J. Higgins and I. W. Dawes, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 6564–6569.
27 A. P. Gerber, D. Herschlag and P. O. Brown, *PLoS Biol.*, 2004, **2**, e79.
28 J.-P. Reichheld, T. Vernoux, F. Lardon, M. Van Montagu and D. Inze, *Plant J.*, 1999, **17**, 647–656.
29 M. G. Koerkamp, M. Rep, H. J. Bussemaker, G. P. M. A. Hardy, A. Mul, K. Piekarska, C. A.-K. Szigyarto, J. M. T. de Mattos and H. F. Tabak, *Mol. Biol. Cell*, 2002, **13**, 2783–2794.
30 Y.-M. Zhang, J.-K. Liu and T.-Y. Wong, *Mol. Microbiol.*, 2003, **48**, 1317–1323.
31 M. A. Singer and S. Lindquist, *Mol. Cell*, 1998, **1**, 639–648.
32 J. Parrou, M. Teste and J. Francois, *Microbiology (Reading, U. K.)*, 1997, **143**, 1891–1900.
33 N. J. Haughey, R. G. Cutler, A. Tamara, J. C. McArthur, D. L. Vargas, C. A. Pardo, J. Turchan, A. Nath and M. P. Mattson, *Ann. Neurol.*, 2004, **55**, 257–267.
34 K. Nakayama, T. Yamaguchi, T. Doi, Y. Usuki, M. Taniguchi and T. Tanaka, *J. Biosci. Bioeng.*, 2002, **94**, 207–211.
35 P. P. Ruvolo, *Leukemia*, 2001, **15**, 1153–1160.
36 S. A. Summers, *Prog. Lipid Res.*, 2005, **45**, 42–72.
37 R. Cutler, J. Kelly, K. Storie, W. Pedersen, A. Tammara, K. Hatanpaa, J. Troncoso and M. Mattson, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 2070–2075.
38 G. Daum, N. D. Lees, M. Bard and R. Dickson, *Yeast*, 1998, **14**, 1471–1510.
39 S. M. Mandala, R. Thornton, Z. Tu, M. B. Kurtz, J. Nickels, J. Broach, R. Menzeleev and S. Spiegel, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 150–155.
40 L. L. Listenberger, D. S. Ory and J. E. Schaffer, *J. Biol. Chem.*, 2001, **276**, 14890–14895.
41 G. Han, K. Gable, S. D. Kohlwein, F. Beaudoin, J. A. Napier and T. M. Dunn, *J. Biol. Chem.*, 2002, **277**, 35440–35449.
42 Y.-A. Moon and J. D. Horton, *J. Biol. Chem.*, 2003, **278**, 7335–7343.
43 V. Luu-The, P. Tremblay and F. Labrie, *Mol. Endocrinol.*, 2006, **20**, 437–443.
44 D. Song, G. Liu, V. Luu-The, D. Zhao, L. Wang, H. Zhang, G. Xueling, S. Li, L. Desy, F. Labrie and G. Pelletier, *J. Steroid Biochem. Mol. Biol.*, 2006, **101**, 136–144.
45 M. Wickens, D. S. Bernstein, J. Kimble and R. Parker, *Trends Genet.*, 2002, **18**, 150–157.
46 F. L. Moore, J. Jaruzelska, M. S. Fox, J. Urano, M. T. Firpo, P. J. Turek, D. M. Dorfman and R. A. R. Pera, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 538–543.
47 I. Ulitsky and R. Shamir, *BMC Syst. Biol.*, 2007, **1**, 8.
48 S. Zhang, X. Ning and X.-S. Zhang, *Comput. Biol. Chem.*, 2006, **30**, 445–451.
49 P. V. Gopalacharyulu, E. Lindfors, J. Miettinen, C. Bounsaythip and M. Oresic, *Int. J. Data Mining Bioinformatics*, 2008, **2**, 54–77.
50 I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *Nucleic Acids Res.*, 2000, **28**, 289–291.
51 T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. Pilar Garcia Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu and R. Apweiler, *Nucleic Acids Res.*, 2006, **35**, D16–D20.
52 M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, *Nucleic Acids Res.*, 2004, **32**, D277–280.
53 V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender, *Nucleic Acids Res.*, 2003, **31**, 374–378.
54 N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann and M. Gerstein, *Nature*, 2004, **431**, 308–312.
55 The UniProt Consortium, *Nucleic Acids Res.*, 2007, **35**, D193–197.
56 M. Katajamaa, J. Miettinen and M. Oresic, *Bioinformatics*, 2006, **22**, 634–636.
57 R. Nash, S. Weng, B. Hitz, R. Balakrishnan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, M. S. Livstone, R. Oughtred, J. Park, M. Skrzypek, C. L. Theesfeld, G. Binkley, Q. Dong, C. Lane, S. Miyasato, A. Sethuraman, M. Schroeder, K. Dolinski, D. Botstein and J. M. Cherry, *Nucleic Acids Res.*, 2007, **35**, D468–471.