

Network motifs that recur across species, including gene regulatory and protein–protein interaction networks

Robert Borotkanics · Harold Lehmann

Received: 18 February 2014 / Accepted: 13 May 2014 / Published online: 22 May 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract Cellular molecules interact in complex ways, giving rise to a cell's functional outcomes. Conscientious efforts have been made in recent years to better characterize these patterns of interactions. It has been learned that many of these interactions can be represented abstractly as a network and within a network there in many instances are network motifs. Network motifs are subgraphs that are statistically overrepresented within networks. To date, specific network motifs have been experimentally identified across various species and also within specific, intracellular networks; however, motifs that recur across species and major network types have not been systematically characterized. We reason that recurring network motifs could potentially have important implications and applications for toxicology and, in particular, toxicity testing. Therefore, the goal of this study was to determine the set of intracellular, network motifs found to recur across species of both gene regulatory and protein–protein interaction networks. We report the recurrence of 13 intracellular, network motifs across species. Ten recurring motifs were found across both protein–protein interaction networks and gene regulatory networks. The significant pair motif was found to recur only in gene regulatory networks. The diamond and one-way cycle reversible step motifs were found to recur only in protein–protein interaction networks. This study is the first formal review of recurring, intracellular network motifs across species. Within toxicology, combining our understanding

of recurring motifs with mechanism and mode of action knowledge could result in more robust and efficient toxicity testing models. We are sure that our results will support research in applying network motifs to toxicity testing.

Keywords Network motif · Subgraph · Cancer · Toxicology

Introduction

Cellular molecules interact in complex ways, giving rise to a cell's functional outcomes. These interactions span an array of functions ranging from glycolysis to transcription to apoptosis. Scientists and researchers have produced volumes of information on how these interactions occur, often describing experimentally the interaction between two or more molecules and the associated mechanism of action. While the interactions between individual molecules and in some instances larger pathways have been characterized to varying degrees, this understanding is variable and often quite incomplete. System biology has in recent years made a conscientious effort to better characterize these broader patterns of interactions.

These complex molecular interactions can be represented abstractly as a network. Biological networks exhibit organizing principles, and further, this organization may be represented by the network's topology (Albert and Barabasi 2002). This organization is distinct from random networks and is very often scale free, which means that their degree distribution approximates a power law, $P(k) \sim k^{-\gamma}$ (Barabasi and Albert 1999; Barabasi and Oltvai 2004).

Mathematically, a cellular network may be represented as a graph (Kaveh 2013), which abstractly comprises nodes (vertices) and edges (arcs). Cellular molecules may be

R. Borotkanics (✉)
Department of Environmental Health Sciences, Johns Hopkins
Bloomberg School of Public Health, Baltimore, MD, USA
e-mail: rborotk1@jhu.edu

H. Lehmann
Division of Health Sciences Informatics, Johns Hopkins School
of Medicine, Baltimore, MD, USA

represented as nodes. Nodes may be connected by an edge—directed or undirected, representing the interactions between the molecules. If the nodes and edges are countable, then the graph is said to be finite. One may further break down a graph into subgraphs. These subgraphs are made up of a finite number of nodes and these nodes' interactions with one another characterize a given network at the local level. Two or more graphs or subgraphs are considered isomorphic if they have the same node number and their adjacency is preserved (i.e., corresponding nodes are connected by edges the same way in both subgraphs). Subgraphs that are statistically overrepresented in a network are referred to as network motifs¹ (Milo et al. 2002; Shen-Orr et al. 2002). Milo et al. (2002) and Shen-Orr et al. (2002) were the first to use graph theory principles and the experimental method to identify motifs within the transcriptional networks of simple organisms. Since their discovery, much research has been carried out identifying and characterizing motifs at the intracellular level across many species.

However, motifs that recur across species have not been systematically characterized to date. We reason that motifs that recur across species could potentially have important implications and applications. Recurring motifs could improve upon our existing knowledge of toxicology and, in particular, could have practical applications for toxicity testing. The US National Academies of Sciences (NAS) in 2007, via its report entitled, *Toxicity Testing in the 21st Century*, posited a vision for toxicity testing, asserting that the goals of toxicity testing moving forward should be to identify those critical pathways that when sufficiently perturbed lead to adverse health outcomes (Mantus et al. 2007). An important foundation and component of these critical pathways, we reason, is motifs. In particular, it is important from the perspective of toxicology to identify and characterize those motifs that recur across species as such knowledge could enable more robust toxicity tests and better assist in informing cross-species extrapolation. This, we reason, sets a stronger foundation for alternative testing methods, like in vitro and in silico, consistent with the NAS report.

The goal of this study therefore is to determine the set of motifs that have been identified at the intracellular level and are found to recur across species and major types of biological networks (i.e., protein–protein interaction and gene regulation). To date, a finite number of motifs have been identified. Motif identification has been carried out extensively in simple organisms, for instance, the transcription networks in *Escherichia coli* and *Saccharomyces cerevisiae* (Dorbin et al. 2004; Konagurthu and Lesk 2008a, b; Lee et al. 2002; Ma et al. 2004; Mazurie et al. 2005; Milo et al. 2002, 2004; Shen-Orr et al. 2002; Yeger-Logem et al.

2004; Zhang et al. 2005). Select motifs have been summarized to varying degrees in the literature (Alon 2007a; Shoval and Alon 2010). However, these motifs have not been characterized via a formal review or in a systematic manner. Herein, we report the intracellular motifs that recur across species. Further, we reaffirm that select motifs are more or less common within broad networks and across species, with specific exceptions. We finally discuss the potential implications and applicability of these findings with respect to toxicology.

Methods

We carried out a search of the peer-reviewed literature, wherein our strategy was to maximize the chances of identifying all available articles, communicating their findings in English. Peer-reviewed research articles were identified using Google Scholar, PubMed, and Scopus. Articles were found by the use of each search engine via keyword search, using the following search terms: 'functional motif,' 'graph,' 'motif,' 'network motif,' or 'subgraph.' Terms were used in their singular and plural forms. Supplementary searches were carried out using the aforementioned terms in combination with the following terms: 'biological network,' 'cellular biology,' 'computational biology,' 'molecular biology,' 'network,' 'network topology,' or 'systems biology,' again using both singular and plural forms.

Articles were included if they reported on research that quantitatively assessed the presence of motifs at the intracellular level, using experimental methods of non-plant organisms. In circumstances where studies were carried out using cell lines, only those studies using non-carcinogenic cell lines were accepted for inclusion. Studies identifying intercellular motifs were excluded. Summaries and reviews were noted and used for informational purposes, but excluded from the analysis, so as to avoid the redundant reporting of findings.

Pertinent motif data from each study—motif structure, organism, data source, biological network, and methods—were documented. A network motif is a recurring pattern within the network that recurs more often than at random (Alon 2007b). These comparisons are made by evaluating the subject network to reasonably equivalent, randomized networks. These can be described by a number of statistical tests. These can be described by a Z score, which is simply the difference in occurrence of a pattern in a selected network compared to a randomized network, divided by the standard deviation of occurrence in the randomized network. If multiple, randomized networks are used, then the mean frequency of the randomized networks is applied (Schwöbbermeyer 2008). The result of this calculation can alternatively be described as a p value. Statistically

¹ Herein, called 'motifs' for brevity.

significant motifs from within each study were recorded, where a motif exhibited a $p < 0.05$, $z > 2$ or a normalized Z score > 0.5 compared to a randomized network or networks. Motif significance profiles are a vector of Z scores of a set of motifs. These allow for comparison of networks of different sizes (Schwöbbermeyer 2008), as was the case in some studies evaluated. Similarly, if a motif exhibited a positive significance profile (SP), it too was accepted for inclusion. The statistically significant motifs from a given study were then compared to the statistically significant motifs identified across all other studies where statistically significant motifs were also identified. A motif was considered to recur if it was identified in at least two independent studies. The result of this analysis was a set of recurring, statistically significant motifs. These recurring motifs were also associated with multiple organisms and intracellular, molecular networks. The primary author carried out the analysis, of which the secondary author carried out a sequential review.

Results

The literature search resulted in the identification of 141 articles, of which 17 met inclusion criteria (Fig. 1; Table 1). Based on these studies, we report 13 recurring motifs (Tables 2, 3, 4, 5, 6). Each edge is directional, but is non-specific with regard to type (i.e., induction, inhibition or binding), unless otherwise stated, consistent with the protocol laid out in the ‘Methods’ section. We organized motif sub-types according to the direction of the higher order motifs’ edges. Of the 13 recurring motifs, there were two sub-types also found to be statistically significant across two or more studies.

Most motifs were found to recur across an evolutionary range of organisms. For instance, the feed-forward loop motif was found to recur in organisms ranging from *S. cerevisiae* to *H. sapiens* and 10 other organisms (Table 3). Only two motifs did not recur across an evolutionary range of organisms: the mixed feedback loop and diamond motifs. Most motifs were found to recur in both the broad categories of gene regulation and protein–protein interaction (PPI). There were, however, exceptions. The one-way cycle reversible step and diamond motifs were found only in PPI networks. The significant pair motif was found solely in gene regulatory networks.

Two studies included in this analysis evaluated motifs where gene regulation and protein–protein interaction were considered jointly as a single network (Yeger-Logem et al. 2004; Zhang et al. 2005). Both studies identified the same, 3-node motifs in *S. cerevisiae*, including the following motifs: feed-forward loop, co-regulator, co-regulated interacting, and protein clique.

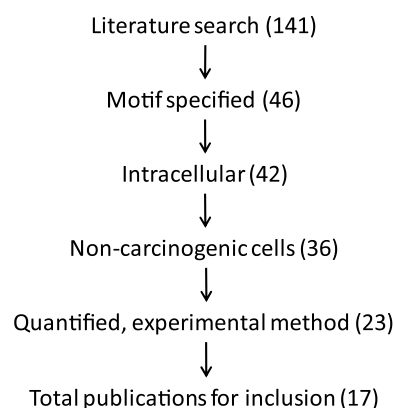


Fig. 1 Inclusion criteria

Sub-networks, or pathways, of gene regulatory and PPI networks were not evaluated, on account of the limited number of studies that evaluated any given pathway.

Methodologically, studies used wide-ranging, overlapping and distinct data sets. We evaluated the effect the choice of data set would have on results, comparing the motifs found in the gene regulatory network of *S. cerevisiae*. Six studies evaluated motifs based on this condition, of which all six experimentally identified the feed-forward loop. Two studies, Joshi et al. (2011) and Mazurie et al. (2005) identified the significant pair motif. Four other studies individually identified the two-way interaction, co-regulator, 3-cycle, and bifan motifs. These four motifs were found to recur only in the larger analysis, which included all organisms and/or networks/pathways studied. This finding indicates that the choice of data set in motif identification analyses is important. It is based on this finding that we reason that it is crucial to continue to conduct motif identification analyses across data sets and spanning a range of organisms, so as to minimize the risk of spurious results.

The majority of studies included in this review applied the isomorphism and randomization methods first developed by Milo et al. 2002 and Kashtan et al. 2004, with minor variations ($n = 13$). There exist a number of distinct, motif isomorphism and randomization algorithms (Wong et al. 2012). Further, the choice of isomorphism algorithm is known to effect results (Shreiber and Schwöbbermeyer 2005), which too was reaffirmed by analysis of the studies included in this review. For instance, Konagurthu and Lesk (2008a, b) repeated the studies of Ma’ayan et al. (2005) and Shen-Orr et al. (2002), but used a more conservative isomorphism algorithm. This research team identified both the feed-forward loop and 3-cycle motifs, consistent with the other two studies. However, Konagurthu and Lesk did not identify the same extent of three-node motifs found in Ma’ayan or the four node motifs of both Ma’ayan et al. and Shen-Orr et al. Further, Kashtan et al. (2004) applied

Table 1 Studies meeting inclusion criteria

Study	Organism/cell line	Data source	Network; pathway (where applicable)	Motif identification method (isomorphism; randomization)
Eom et al. (2006)	<i>S. cerevisiae</i>	Alon (provided)	Gene regulation; transcription	Milo et al. (2002); edge swapping
Joshi et al. (2011)	<i>S. cerevisiae</i>	(1) Halbeisen and Gerber (2009); (2) Harbison et al. (2004)	Gene regulation; posttranscriptional and translational control of mRNA	FANMOD (Wernicke and Rasche 2006); unstated
Kashani et al. (2009)	<i>E. coli</i>	KEGG	PPI; metabolism	Demonstrates algorithm based on revolving door ordering and NAUTY; Milo et al. (2002)
Kashtan et al. (2004)	<i>E. coli</i>	Shen-Orr et al. (2002) (data v. 1.1)	Gene regulation; transcription	Demonstrates an algorithm for motif identification based on sampling; switching, matching, go with winners
Kim et al. (2011)	<i>D. melanogaster</i> <i>E. lupus</i> <i>H. sapiens</i> <i>M. musculus</i> <i>R. norvegicus</i> <i>S. cerevisiae</i>	Nikitin et al. (2003)	PPI	Milo et al. (2002); edge swapping
Konagurthu and Lesk (2008a, b)	<i>E. coli</i> Mammalian hippocampal CA1 neuron	Shen-Orr et al. (2002) (data v 1.0) Ma'ayan et al. (2005)	Gene regulation; transcription PPI: signal propagation resulting from ligand occupancy	Ullmann (1976); Shen-Orr et al. (2002) for 3 node randomization. For N node randomization: Konagurthu and Lesk (2008a, b)
Ma'ayan et al. (2005)	<i>S. cerevisiae</i> Mammalian hippocampal CA1 neuron	Luscombe et al. (2004) Literature search	Gene regulation; transcription PPI: signal propagation resulting from ligand occupancy	Kashtan et al. (2004): full enumeration (3–4 nodes). Random sampling (5–6 node); unstated
Martinez et al. (2008)	<i>C. elegans</i>	miRNA- Yeast 1-hybrid method	Gene regulation: transcription of miRNAs and posttranscriptional control of transcription Factors by miRNAs	Kashtan et al. 2004; edge switching, node replacement, complete randomization
Mazurie et al. (2005)	<i>S. cerevisiae</i>	(1) Guelzim et al. (2002) (2) Database of interacting proteins	Gene regulation: transcription	Milo et al. (2002); Maslov and Snippen (2002)
Milo et al. (2002)	<i>E. coli</i> <i>S. cerevisiae</i>	Shen-Orr et al. (2002) Yeast proteome database	Gene regulation: transcription Gene regulation: transcription	Developed method; randomization by 2 algorithms: (1) Markov, based on Kannan et al. (1997), Maslov and Snippen (2002); (2) modification of Newman et al. (2001)

Table 1 continued

Study	Organism/cell line	Data source	Network; pathway (where applicable)	Motif identification method (isomorphism; randomization)
Milo et al. (2004)	<i>B. Subtilis</i> <i>D. melanogaster</i> <i>E. coli</i> <i>E. perischoeichinoidea</i> <i>H. sapiens</i> <i>S. cerevisiae</i>	Ishii et al. (2001) GeneNet Shen-Orr et al. (2002) Davidson et al. (2002) Signal transduction knowledge environment Milo et al. (2002) and Lee et al. (2002)	Gene regulation: transcription Gene regulation: developmental transcription networks Gene regulation: transcription Gene regulation: endomesoderm development transcription networks PPI: signal transduction Gene regulation: transcription	Mfinder 1.1 (Milo et al. 2002; Kashtan et al. 2004); edge switching
Ryall et al. (2012) Shalgi et al. (2007)	Neonatal <i>R. norvegicus</i> myocytes <i>E. lupus</i> <i>H. sapiens</i> <i>M. musculus</i> <i>R. norvegicus</i> <i>H. salinarum</i> <i>H. sapiens</i> <i>M. acetivorans</i> <i>M. barkeri</i> <i>M. musculus</i> <i>S. cerevisiae</i> <i>E. coli</i>	Literature search (1) TargetScan; (2) PicTar Gonzalez et al. (2008) Duarte et al. (2007) Kumar et al. (2011) Feist et al. (2006) Sigurdsson et al. (2010) Duarte et al. (2004) (1) Regulon Database (v3.2); (2) Literature search	PPI: cardiac hypertrophy Gene regulation: transcription of miRNAs and posttranscriptional control of transcription factors by miRNAs PPI: metabolism Gene regulation: transcription	NetMatch; RandomNetworks Shen-Orr et al. (2002); edge swapping
Shellman et al. (2013)	<i>H. sapiens</i> <i>M. musculus</i> <i>S. cerevisiae</i> <i>E. coli</i>	Gonzalez et al. (2008) Duarte et al. (2007) Kumar et al. (2011) Feist et al. (2006) Sigurdsson et al. (2010) Duarte et al. (2004) (1) Regulon Database (v3.2); (2) Literature search	PPI: metabolism Gene regulation: transcription	FANMOD (Wernicke and Rasche 2006); unstated
Shen-Orr et al. (2002)	<i>E. coli</i>	(1) Regulon Database (v3.2); (2) Literature search	Gene regulation: transcription	Connectivity matrix for non-dense overlapping regions and standard average-linkage algorithm for dense overlapping regions; Newman et al. (2001) and Kannan et al. (1997)
Yeager-Logem et al. (2004)	<i>S. cerevisiae</i>	(1) Biomolecular interaction network database; (2) Database of interacting proteins; (3) Lee et al. (2002); (4) Munich information center for protein sequences database; (5) Ren et al. (2000); (6) SCPD promoter database; (7) Simon et al. (2001); (8) Yeast proteome database	Combined gene regulation and PPI	Extension of Shen-Orr et al. (2002), considering extended degree of a node and the edge profile of two nodes; unstated

Table 1 continued

Study	Organism/cell line	Data source	Network; pathway (where applicable)	Motif identification method (isomorphism; randomization)
Zhang et al. (2005)	<i>S. cerevisiae</i>	(1) Gavin et al. (2002); (2) Ho et al. (2002); (3) Hughes et al. (2000); (4) Lee et al. (2002); (5) Munich information center for protein sequences database; (6) Protein sequence homology; relationships from a genome-wide BLAST search; (7) Tong et al. (2004)	Combined gene regulation and PPI	Milo et al. (2002); Park and Newman (2003)

Table summarizes only those organisms and pathways the studies reported statistical results on

a sampling-based algorithm to a version of the data used in Shen-Orr et al. (2002), identifying feed-forward loops and dense overlapping regions, but like Konagurthu and Lesk, did not make a single input module motif finding.

Only seven studies provided explicit motif definitions. Therefore, in many instances, many terms and definitions in Tables 2, 3, 4, 5 and 6 had to be inferred based on the larger study description and associated diagrams. The unique term to identify a motif type in this analysis was generally that term used by the research team that made the initial finding. In limited instances, more refined or concise terms were applied for sake of clarity. For instance, Shen-Orr et al. (2002) were the first to use *feed-forward loop*, and so is applied (Table 3). We applied the same rule to the definitions of each motif. We did abstract each definition, so that it could be applied to describe such motifs regardless of network type in those instances where motifs were found across both network types, with the goal of affording the motif adequate, generalized description. For instance, Shen-Orr et al. (2002) first defined the FFL as, *defined by a transcription factor X that regulates a second transcription factor Y, such that both X and Y jointly regulate an operon Z*. As FFLs are known to reside in gene regulatory and PPI networks, we abstracted this definition to, *A regulates B, such that both A and B jointly regulate an operon C*. We hope this approach to terms and definitions minimizes controversy.

Finally, a limited set of motifs could not be curated. This is because these motifs were found in only single studies or lacked complete information.

Discussion

These 13 recurring motifs suggest that they are more than just statistical anomalies; they potentially represent critical patterns of biological activity essential for cellular homeostasis. This study constitutes the first formal review of recurring molecular motifs across species and pathways at the intracellular level. The review was based on primary research that drew conclusions originating from distinct research teams, unique information sources, using overlapping methods.

Other studies generally support this review's findings. Alon (2007a) reported that feed-forward loops recur in the transcription networks that respond to environmental stimuli of *E. coli*, *H. sapiens* and *S. cerevisiae*. The same summary reported on the recurrence of dense overlapping regions in the transcription networks in *E. coli* and *S. cerevisiae*. Eom et al. (2006) carried out a taxonomic analysis of motif triads in energy metabolism pathways across 43 species and found many motifs were largely conserved within broad taxonomic categories.

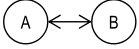
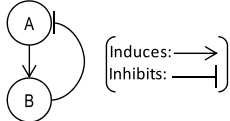
Table 2 Two-node motifs

Motif	Two-way interaction	Mixed feedback loop
Definition	A regulates the activity of or the production of B. B or the product of B regulates the activity of or the production of A.	A regulates the production of B. The product of B inhibits the activity of A.
Pathways	GR: transcription of miRNAs & post transcriptional control of transcription factors by miRNAs; PPI	GR: transcription of miRNAs & post transcriptional control of TFs by miRNAs; Combined transcription-protein-protein interaction
Organisms	<i>D. melanogaster</i> ; <i>E. lupus</i> ; <i>H. sapiens</i> ; <i>M. musculus</i> ; <i>R. norvegicus</i> ; <i>S. cerevisiae</i>	<i>C. elegans</i> ; <i>S. cerevisiae</i>
References	Kim <i>et al.</i> , 2011; Mazurie <i>et al.</i> , 2005	Martinez <i>et al.</i> , 2008; Yeager-Lotem <i>et al.</i> , 2004

Legend:

Gene regulation: GR

Protein-protein interaction: PPI

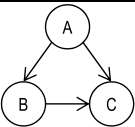
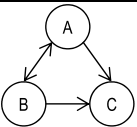
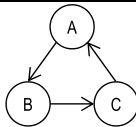
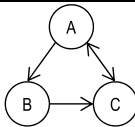



(Induces: →)

(Inhibits: —|)

Table 3 Three-node motifs (1 of 2)

Motif	Feed Forward Loop	Co-regulator	3-Cycle	One way cycle, reversible step
Definition	A regulates B, such that both A and B jointly regulate an operon C.	Two regulators, A & B that interact with one another, both jointly regulating the activity or production of the target, C.	A regulates the activity or production of B, B regulates the activity or production of C and C regulates the activity or production of A.	A regulates the activity of B and C, B regulates the activity of C, C regulates the activity of A.
Pathways	GR: post transcriptional control of transcription factors by miRNAs, transcription; PPI: cardiac hypertrophy, metabolism, signal transduction	GR: post transcriptional & translational control of miRNA, transcription; PPI: signal propagation resulting from ligand occupancy	GR: transcription, transcription in response to stress; PPI: signal propagation resulting from ligand occupancy	PPI: metabolism
Organisms	<i>B. subtilis</i> ; <i>D. melanogaster</i> ; <i>E. coli</i> ; <i>E. lupus</i> ; <i>E. perischoechinoidea</i> ; <i>H. salinarum</i> ; <i>H. sapiens</i> ; Mammalian hippocampal CA1 neuron; <i>M. acetivorans</i> ; <i>M. bakeri</i> ; <i>M. musculus</i> ; <i>R. norvegicus</i> ; Neonatal <i>R. norvegicus</i> myocytes; <i>S. cerevisiae</i>	<i>D. melanogaster</i> ; <i>E. lupus</i> ; <i>H. sapiens</i> ; Mammalian hippocampal CA1 neuron; <i>M. musculus</i> ; <i>R. norvegicus</i> ; <i>S. cerevisiae</i>	<i>E. coli</i> ; Mammalian hippocampal CA1 neuron; <i>S. cerevisiae</i>	<i>H. salinarum</i> ; <i>H. Sapiens</i> ; <i>M. musculus</i> ; <i>R. norvegicus</i> ; <i>S. cerevisiae</i>
References	Eom <i>et al.</i> , 2012; Joshi <i>et al.</i> , 2011; Kashtan <i>et al.</i> , 2004; Kim <i>et al.</i> , 2011; Konagurthu & Lesk 2008; Mazurie <i>et al.</i> , 2005; Milo <i>et al.</i> , 2002; Milo <i>et al.</i> , 2004; Ryall <i>et al.</i> , 2012; Shalgi <i>et al.</i> , 2007; Shellman <i>et al.</i> , 2013; Shen-Orr <i>et al.</i> , 2002; Yeager-Lotem <i>et al.</i> , 2004; Zhang <i>et al.</i> , 2005	Joshi <i>et al.</i> , 2011; Kim <i>et al.</i> , 2011; Ma'ayan <i>et al.</i> , 2005; Shalgi <i>et al.</i> , 2007; Yeager-Lotem <i>et al.</i> , 2004; Zhang <i>et al.</i> , 2005	Kim <i>et al.</i> , 2011; Konagurthu & Lesk 2008; Ma'ayan <i>et al.</i> , 2005	Kim <i>et al.</i> , 2011; Shellman <i>et al.</i> , 2013

(Induces: →)

(Binds: —●)

While these findings reinforce this study's strength, there too are study limitations. For instance, it could be that motifs simplify a multistep processes into over-simplified interactions and so could under or over-state the existence of a motif non-differentially. This could be due to the fact that the knowledge of a given network or pathway may not be complete or the methods used to integrate network data from multiple sources are unclear or not specified to the right level of detail. Second, potentially important motifs that do not recur would not be detected based on the general methods used to inform this review. Third, this study

applied conservative criteria in accepting a motif as recurring. Further, most studies did not specify the type of relationship (edge) between nodes (i.e., induction, inhibition or binding). Only those studies published in English were included. The results of this study, therefore, most likely under represent the full catalog of recurring motifs and subtypes across species and networks.

These limitations are largely overcome by the overall strengths of this review. This review includes analyses from 17 studies. These studies experimentally investigated the existence of motifs within sub-cellular networks ranging

Table 4 Three-node motifs (2 of 2)

Motif	Co-regulated Interacting	Protein Clique	Significant Pair
Definition	B and C are regulated by a common regulator, A, B and C and their products interact with each other.	A, B and C join, forming a complex.	A and B regulate the production of C.
Pathways	GR: transcription; PPI: signal propagation resulting from ligand occupancy; signal transduction	GR: transcription; PPI: metabolism, signal propagation resulting from ligand occupancy, signal transduction	GR: post transcriptional & translational control of miRNA, transcription
Organisms	<i>H. sapiens</i> ; Mammalian hippocampal CA1 neuron; <i>M. musculus</i> ; <i>R. norvegicus</i> ; <i>S. cerevisiae</i>	<i>E. coli</i> ; <i>H. sapiens</i> ; <i>M. bakeri</i> ; Mammalian hippocampal CA1 neuron; <i>M. musculus</i> ; <i>R. norvegicus</i> ; <i>S. cerevisiae</i>	<i>S. cerevisiae</i>
References	Kim <i>et al.</i> , 2011; Ma'ayan <i>et al.</i> , 2005; Milo <i>et al.</i> , 2004; Yeger-Lotem <i>et al.</i> , 2004; Zhang <i>et al.</i> , 2005	Kim <i>et al.</i> , 2011; Ma'ayan <i>et al.</i> , 2005; Milo <i>et al.</i> , 2004; Shellman <i>et al.</i> , 2013; Yeger-Lotem <i>et al.</i> , 2004; Zhang <i>et al.</i> , 2005	Joshi <i>et al.</i> , 2011; Mazurie <i>et al.</i> , 2005

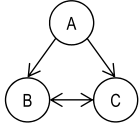
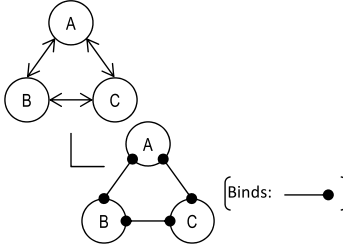
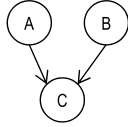




Table 5 Four-node motifs

Motif	Bifan	Diamond
Definition	Two upstream components, A & B, independently or co-regulate the activity or production of two downstream components, C & D.	A regulates the activity of B & C. C & D regulate the activity of D.
Pathways	GR: transcription; PPI: cardiac hypertrophy, signal propagation resulting from ligand occupancy	PPI: cardiac hypertrophy, signal propagation resulting from ligand occupancy
Organisms	<i>E. coli</i> ; mammalian hippocampal CA1 neuron; Neonatal <i>R. norvegicus</i> myocytes; <i>S. cerevisiae</i>	Mammalian hippocampal CA1 neuron; Neonatal <i>R. norvegicus</i> myocytes
References	Kashlan <i>et al.</i> , 2004; Ma'ayan <i>et al.</i> , 2005; Milo <i>et al.</i> , 2002; Ryall <i>et al.</i> , 2012; Yeger-Lotem <i>et al.</i> , 2004	Ma'ayan2005; Ryall2012

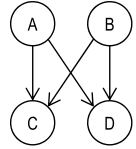
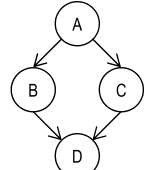
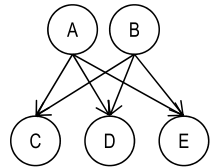
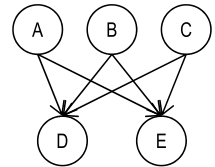



Table 6 Five-node motif

Motif	Dense Overlapping Region 1	Dense Overlapping Region 2
Definition	Is an overlapping layer of regulators, A & B, independently or co-regulating the production or activity of operons, C, D & E.	Is an overlapping layer of regulators, A, B & C, independently or co-regulating the production or activity of operons, D & E.
Pathways	GR: transcription; PPI: signal propagation resulting from ligand occupancy	GR: transcription; PPI: signal propagation resulting from ligand occupancy
Organisms	<i>E. coli</i> ; Mammalian hippocampal CA1 neuron	<i>E. coli</i> ; Mammalian hippocampal CA1 neuron
References	Kashtan <i>et al.</i> , 2004; Ma'ayan <i>et al.</i> , 2005; Shen-Orr <i>et al.</i> , 2002	Kashtan <i>et al.</i> , 2004; Ma'ayan <i>et al.</i> , 2005

from simple organisms, like *S. cerevisiae*, to *H. sapiens*. Further, these motifs were found consistently regardless of data source or research method and across distinct research teams.

It is important to note the absence of three subgraphs from the recurring set: (1) auto-regulation; (2) cascade; and (3) single input module (SIM). Auto-regulation is where a molecule directly induces or inhibits the activity or production of itself. Auto-regulatory subgraphs are known to occur with some transcription factors, where the feedback edge is often inhibitory; e.g., Nrf2² exhibits this pattern. The cascade subgraph is where molecule A regulates the activity of B; B regulates the activity of C, as demonstrated in the MAPK³ cascade. The SIM is defined by a set of operons that are controlled by a single transcription factor (Shen-Orr et al. 2002). Aryl hydrocarbon receptor, for example, is known to induce the transcription of a multitude of mRNA.

Each of these subgraphs was identified experimentally in single studies; however, none were experimentally identified in two or more studies using the inclusion criteria and in subgraphs of the same node number. The absence of auto-regulation may be an artifact. Many studies excluded auto-regulatory feedback loops from their network graphs before carrying out analyses. Further, most studies carried out analyses only at subgraphs of node size three or greater. Such study designs would preclude identification of auto-regulatory loops in most circumstances. The cascade subgraphs, may too be underrepresented, but for a different reason. Joshi et al. (2011) identified a three-node cascade subgraph in the posttranscriptional control of mRNA of *S. cerevisiae*. Kashani et al. (2009) identified a four node form of the same subgraph in the metabolic network of *E. coli*. We reason that underrepresentation of the cascade subgraph exist, in part because most studies evaluated pathways associated with the early steps of gene regulation, and few studies researched the complete gene regulation pathway. The regulatory properties of transcription factors and associated molecules typically do not exhibit cascade-like topology. Further, signaling pathways, which are generally regarded to exhibit cascade-like topology, have not been as extensively researched as have transcription networks. Of the studies included in this systematic review, only seven evaluated PPIs.

Finally, the SIM subgraph was experimentally identified in only a single study (Shen-Orr et al. 2002). We reason based on our own experiences that this subgraph is likely a recurring motif. We evaluated potential causes of underrepresentation, but analysis of the studies used in this review did not reveal one single cause. Choice of isomorphism algorithm does impact results. Further, selected pathways, data sets, and definitions may also be contributors. Clear

articulation of these important methodological factors is critical in motif identification studies.

Four studies did research subgraphs of node sizes five or greater (Kashani et al. 2009; Kashtan et al. 2004; Ma'ayan et al. 2005; Shen-Orr et al. 2002). The five-node, dense overlapping region (DORs) motif was found across three studies, as noted in the 'Results' section. Other five-node or larger-sized motifs were not found to recur. The reason for this may in part be due to the fact that the computational complexity associated with detecting larger motifs is expensive, and it has only been in recent years that the ability to conduct such analyses at this level has become available.

Motifs that may exist within specific biological pathways, particularly in *H. sapiens* compared to other species, represent areas where further research is needed. Related, analytical tools now exist where specific motif sub-types may be identified. For instance, researchers may now use color techniques to discern motifs of the same node size with specific types of nodes and edges (e.g., FANMOD or similar tools). This is an important analytical capability for motif identification in more complex pathways like signaling or pathways involved in adaptation to environmental stimuli, as these pathways exhibit complex patterns of induction, inhibition, and binding.

We reason that these recurring motifs could have important implications. For instance, recurring motifs could bring increased knowledge is toxicology and, in particular, toxicity testing. The US National Academies of Sciences (NAS) in 2007, via its report entitled, *Toxicity Testing in the 21st Century*, posited a vision for toxicity testing, asserting that the goals of toxicity testing moving forward should be to identify those critical pathways that when sufficiently perturbed lead to adverse health outcomes (Mantus et al. 2007). The NAS further stated that, 'perturbations of cell-signaling motifs... are obligatory changes related to chemical exposure that might eventually result in disease'. This new paradigm is a sharp contrast to the reductionist tradition of toxicology. However, the rich literature on mechanism and mode of action could be a potentially important source to build upon and apply to our existing knowledge of recurring motifs. For instance, two nuclear receptors, constitutive androstane receptor (CAR), and pregnane X receptor (PXR) are known to regulate the transcription of a common number of phase I and phase II metabolizing enzymes (Maglich et al. 2002). Nuclear receptors are a superfamily of transcription factors and are commonly associated with apical outcomes of interests to toxicologists involved in regulatory decision making (Shah et al. 2011). As we see in Fig. 2, the simplified relationship between CAR and PXR resembles a bipartite graph and a potential Bifan or DOR motif. The DOR motif is poorly understood and simplified dynamic models based on structure alone yield seemingly contradictory results (Ingram

² NF-E2-related factor 2.

³ Mitogen-activated protein kinase.

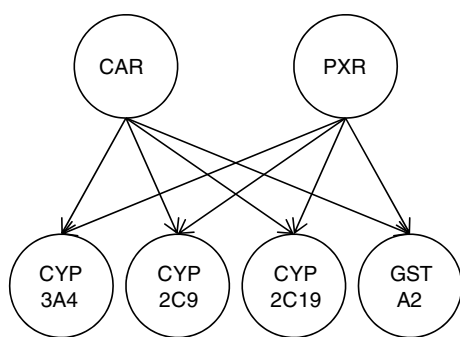


Fig. 2 CAR-PXR DOR

et al. 2006; Alon 2007a). However, in the instance of CAR and PXR, it is known that PXR is more promiscuous and able to recognize a wider range of xenobiotics than CAR, whereas CAR is expressed at higher basal levels (Moore et al. 2003). Further, nuclear receptors are known to undergo posttranscriptional regulation (Takagi et al. 2008). This toxicological information has the potential to better inform dynamic models of recurring motifs and in turn be applied to analyze the effects of chemical perturbations on key components of regulatory pathways, i.e., recurring motifs. This combined approach has the potential to result in validated models, better informing more rapid commercial and regulatory toxicology decisions. Practically, this methodological approach has the potential to reduce reliance on whole animal testing, which is a cost burden to both regulators and industry (Lilienblum et al. 2008).

This first formal review of network motifs finds the recurrence of 13 motifs across species. These 13 recurring motifs are more than just statistical anomalies; they potentially represent critical patterns of biological activity essential for cellular homeostasis and are an important, scientific step forward in further understanding the wiring of sub-cellular pathways. While further research is required, these findings also have potentially important and practical implications for toxicology. We encourage increased applied research to analyze how this knowledge can be applied to chemical safety.

Acknowledgments Mary Fox, DrPH, Johns Hopkins Bloomberg School of Public Health; Michael Trush, PhD, Johns Hopkins Bloomberg School of Public Health; Louis Scarano, PhD, US Environmental Protection Agency.

References

- Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Alon U (2007a) An introduction to systems biology: design principles of biological circuits. Taylor and Francis, Boca Raton
- Alon U (2007b) Network motifs: theory and experimental approaches. *Nat Genet* 8:450–461
- Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
- Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Genet* 5:101–113
- Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan ZJ, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H (2002) A genomic regulatory network for development. *Science* 295(5560):1669–1678
- Dorbin R, Beg QK, Barabasi AL, Oltvai ZN (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinform* 5
- Duarte NC, Herrgård MJ, Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 14:1298–1309
- Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci* 104(6):1777–1782
- Eom Y, Lee S, Jeong H (2006) Exploring local structural organization of metabolic networks using subgraph patterns. *J Theor Biol* 241:823–829
- Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol Syst Biol* 2(2006):0004
- Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147
- Gonzalez O, Gronau S, Falb M, Pfeiffer F, Mendoza E, Zimmer B, Oesterhelta D (2008) Reconstruction, modeling and analysis of *Halobacterium salinarum* R-1 metabolism. *Mol Biosyst* 4:148–159
- Guelzim N, Bottani S, Bourgine P, Képès F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31:60–63
- Halbeisen RE, Gerber AP (2009) Stress-dependent coordination of transcriptome and translatome in yeast. *PLoS Biol* 7(5):e1000105
- Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180–183
- Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD et al (2000) Functional discovery via a compendium of expression profiles. *Cell* 102:109–126
- Ingram PJ, Stumpf MPH, Stark J (2006) Network motifs: structure does not determine function. *BMC Gen* 7(108)
- Ishii T, Yoshida K, Terai G, Fujita Y, Nakai K (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucl Acids Res* 29(1):278–280
- Joshi A, van de Peer Y, Michoel T (2011) Structural and functional organization of RNA regulons in the post-transcriptional regulatory network of yeast. *Nucl Acids Res* 39(21):9108–9117
- Kannen R, Tetali P, Vempala S (1997) Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Struct Algorithms* 14:293

- Kashani ZRM, Ahrabian H, Elahi E, Nowzari-Dalini A, Ansari ES, Asadi S, Mohammadi S, Schreiber F, Masoudi-Nejad A (2009) Kavosh: a new algorithm for finding network motifs. *BMC Bioinform* 10
- Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 20(11):1745–1758
- Kaveh A (2013) Introduction to graph theory and algebraic graph theory. In: *Optimal analysis of structures by concepts of symmetry and regularity*. Springer, New York
- Kim W, Li M, Wang J, Pan Y (2011) Biological network motif detection and evaluation. *BMC Syst Biol* 5(Suppl 3):S5
- Konagurthu AS, Lesk AM (2008a) On the origin of distribution patterns of motifs in biological networks. *BMC Syst Biol* 2(73)
- Konagurthu AS, Lesk AM (2008b) Single and multiple input modules in regulatory networks. *Proteins Struct Funct Bioinform* 73(2):320–324
- Kumar VS, Ferry JG, Marana CD (2011) Metabolic reconstruction of the archaeon methanogen *Methanosarcina acetivorans*. *BMC Syst Biol* 5:28
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne TB, Volkert TL, Reaenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298:799–804
- Lilienblum W, Dekant W, Foth H, Gebel T, Hengstler JG, Kahl R, Kramer PJ, Schweinfurth H, Wollin KM (2008) Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). *Arch Toxicol* 82:211–223
- Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431(7006):308–312
- Ma HW, Kumar B, Dittges U, Gunzer F, Buer J, Zeng AP (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucl Acids Res* 32(22):6643–6649
- Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Keshenbaum A, Stolovitzky GA, Blitzer RD, Lyengar R (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science* 309:1078–1083
- Maglich JM, Stoltz CM, Goodwin B, Hawkins-Brown D, Moore JT, Kliewer SA (2002) Nuclear pregnane X receptor and constitutive androstane receptor regulate overlapping but distinct sets of genes involved in xenobiotic detoxification. *Mol Pharmacol* 62:638–646
- Mantus E, Obernier J, Crossgrove R, Grossblatt N, Karalic-Loncarevic M, Crago J (2007) Toxicity testing in the 21st century; a vision and a strategy. The National Academies Press, Washington, DC
- Martinez NJ, Ow MC, Barrasa IM, Hammell M, Sequerra R, Doucette-Stamm L, Roth FP, Ambros VR, Walhout AJM (2008) A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev* 22:2535–2549
- Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910–913
- Mazurie A, Bottani S, Vergassola M (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol* 6(4)
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827
- Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of evolved and designed networks. *Science* 303:1538–1542
- Moore JT, Moore LB, Maglich JM, Kliewer SA (2003) Functional and structural comparison of PXR and CAR. *Biochim Biophys Acta* 1619:235–238
- Newmann MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64:026118-1–026118-17
- Nikitin A, Sergei E, Nikolai D, Ilya M (2003) Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* 19(16):2155–2157
- Park Y, Newmann MEJ (2003) Origin of degree correlations in the Internet and other networks. *Phys Rev E* 68:026112-1–026112-7
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* 290:2306–2309
- Ryall KA, Holland DO, Delaney KA, Kraeutler MJ, Parker AJ, Sauermaier JJ (2012) Network reconstruction and systems analysis of cardiac myocyte hypertrophy signaling. *J Biol Chem* 287:42259–42268
- Schwöbbermeyer H (2008) Network motifs. In: *Junker B, Schreiber F (eds) Analysis of biological networks*. Wiley, Hoboken
- Shah I, Houck K, Judson RS, Kavloch RJ, Martin MT, Reif DM, Wambaugh J, Dix DJ (2011) Using nuclear receptor activity to stratify hepatocarcinogens. *PLOS One* 6(2):e14584
- Shalgi R, Lieber D, Oren M, Pilpel Y (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput Biol* 3(7):1291–1304
- Shellman ER, Burant CF, Schnell S (2013) Network motifs provide signatures that characterize metabolism. *Mol BioSyst* 9:352–360
- Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31:64–68
- Shoval O, Alon U (2010) SnapShot: network motifs. *Cell* 143:326–326e1
- Shreiber F, Schwöbbermeyer H (2005) Frequency concepts and pattern detection for the analysis of motifs in networks. *Trans Comput Syst Biol* 3:89–104
- Sigurdsson MI, Jamshidi N, Steingrimsdottir E, Thiele I, Palsson BØ (2010) A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol* 4:140
- Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 106:697–708
- Takagi S, Nakajima M, Mohri T, Yokoi T (2008) Post-transcriptional regulation of human pregnane X receptor by micro-RNA affects the expression of cytochrome P450 3A4. *J Biol Chem* 283(15):9674–9680
- Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Beriz GF, Brost RL, Chang M et al (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808–813
- Ullmann JR (1976) An algorithm for subgraph isomorphism. *J Assoc Comput Mach* 23(1):31–42
- Wernicke S, Rasche R (2006) FANMOD: a tool for fast network motif detection. *Bioinformatics* 22(9):1152–1153
- Wong E, Baur B, Quader S, Huang CH (2012) Biological network motif detection: principles and practice. *Brief Bioinform* 13(2):202–205
- Yeger-Logem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *PNAS* 101(16):5939
- Zhang LV, King OD, Wong SL, Goldberg DS, Tong AHY, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* 4(6)