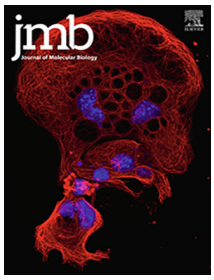
Research Article



Paths Through the Yeast Regulatory Network in Different Physiological States

Arthur M. Lesk [1*] **and Arun S. Konagurthu** [2]

*1 - **Department of Biochemistry and Molecular Biology and Center for Computational Biology and Bioinformatics,** The Pennsylvania State University, University Park, PA 16802, USA*
*2 - **Department of Data Science and Artificial Intelligence,** Monash University, Clayton, VIC 3800, Australia*

Correspondence to Arthur M. Lesk: *aml25@psu.edu (A.M. Lesk)*
https://doi.org/10.1016/j.jmb.2021.167181
Edited by Rita Casadio

Abstract

We analyse paths through the regulatory networks that control gene-expression patterns in Yeast, in five different physiological states: cell cycle, DNA damage, stress response, diauxic shift, and sporulation. The network in each state is specified as a directed graph, containing different sets of edges connecting pairs selected from a combined set of 1475 nodes. Each network contains some nodes that have no parents, and others that have no children. We call these, respectively, 'source' and 'sink' nodes. For each network we enumerate paths between source and sink nodes. In a previous paper (Lesk and Konagurthu, 2020), we defined, extracted and compared the neighbourhoods of each transcription factor in different physiological states, and how the system reconfigures itself. Here we compare the usage of nodes and edges by different networks, and how they are assembled into paths. The picture that emerges is that the networks are not disjoint but show substantial sharing of nodes and edges; however, they assemble these materials into different sets of paths. Four of the networks, other than the cell-cycle network, contain paths between only a small fraction (<13%) of possible source-sink pairs. Although the cell-cycle network is not an outlier in terms of total number of nodes and edges, and number of sink nodes, it is very much an outlier in having a greater proportion of source-to-sink paths than the other networks.

Introduction

Scientists are continuing to learn fascinating details about mechanisms of biological invention. In one classical paradigm, selection adapts structure and function to unique optima defined by external conditions; for instance, beak shape in Darwin's finches. But often the results are not unique. Alternatives may exist side-by-side; for instance, photopic and scotopic vision. More generally, developmental biology deals with internally programmed changes in structure – including metamorphosis, and microscopic correlates in terms of variations in gene expression patterns – that are relatively (although certainly not entirely) free of external influences. But many processes require reactions to changing stimuli: In our daily lives we make a continuous series of conscious and unconscious decisions that adjust our behaviour. At the cellular level such responses take the form of alterations in gene expression patterns.

How does Yeast organise the transcriptional response to changing conditions? We are fortunate to have data sets that describe Yeast under different physiological conditions. There is no reason to think that these exhaust the possible configurations of the network, but they do permit study of some of the possible variety, and detailed description of how it is achieved.

Regulatory networks organise activities within cells, and their interactions with their surroundings. In different physiological states, cells 'reprogram' the regulators of transcription, appropriately to control altered gene-expression patterns. The results confirm the general picture of the elements of the regulatory networks as 'hardware', utilised by the network in different states by assembly of a set of extensively-shared elements (nodes and edges in the networks) into different 'software'.[1]

For yeast, data sets are available that specify transcription control networks in five states: cell cycle, diauxic shift, DNA damage, sporulation, and stress response.[2] Each network is a directed graph. The datasets contain lists of pairs of nodes connected by an edge in the graph. Parent nodes in any edge in any of the networks are *transcription regulators*. Some nodes are *not* parent to any other node in any of the networks; they are regulated but do not regulate other genes. These are *target genes*. Altogether the networks contain 1475 genes, corresponding to approximately half the known proteome of *S. cerevisiae*.

A seminal study of these networks, by Luscombe, et al., reported general statistics of the topologies of the graphs, including comparing networks in different conditions with respect to path lengths, clustering coefficient, indegree and outdegree, and overlap in transcription factor usage.[2] The local structure of the networks includes the canonical 'motifs' described by Milo et al.: the Single-Input Motif (SIM), Multiple-Input Motif (MIM), and Feed-Forward Loop (FFL).[3] Luscombe, et al., and Konagurthu & Lesk studied the distributions of these canonical motifs in the different states of the yeast regulatory network.[2,4,5] Other studies of the networks have focussed on the global structure, and have defined additional local motifs.[6,7] This work extends previous investigations by comparing, in networks corresponding to different physiological states, how transcription factors are assembled into paths.

We are interested in analysing and comparing the large-scale structure of the networks. In a previous paper we studied the nature and distribution of the neighbourhoods of individual nodes.[1] There we defined the *1-neighbourhood* of a node as the subgraph of the network consisting of the selected node, all parent and child nodes of the selected node, plus any additional edges between these parent and child nodes (technically: the vertex-induced subgraph involving the selected node, its parents and its children.) We described the similarities and differences in the structures and topologies of the 1-neighbourhoods of individual nodes in different physiological states, and reported how the networks recombine and reorganise a common set of elements to achieve different purposes.

Here we focus on larger-scale paths through the networks. The goals of the work presented here are:

1. to analyse the usage of nodes and edges in the different networks, and their assembly into paths, and
2. to understand the degree of overlap among the different networks, in terms of the extent of shared nodes, edges and paths.

Each of the five networks contains certain 'source' nodes; that is, nodes that are not the child of any other node, and 'sink' nodes (often called 'target genes') that are not the parent of any other node. The sets of source and sink nodes overlap but are not identical among the five networks. We wrote programs to enumerate, for each network, possible paths between each of the source nodes and each of the sink nodes.

Here we report an analysis of transcription regulatory networks of *Saccharomyces cerevisiae* to compare the assembly of nodes and edges into paths, in the different physiological states.

Results

For each network, we determined the source and sink nodes, and enumerated the paths between each source node and each sink node. In many cases there are multiple paths between a source and a sink node. Figure 1 illustrates an example from the DNA damage network: There are four paths from the source node YDL056W to the sink node YDR523C. One of the paths is short, containing 2 edges (*i.e.,* only one intermediate node between source and sink nodes). The longest path contains 8 edges.

The choice of how to count paths is a matter of considerable nuance. For some pairs of source and sink nodes, it is not possible to enumerate all paths between them, because some paths contain cycles. Repeated traversal of any cycle will generate an infinite number of paths. We have chosen to compute and report the following:

1. For any path between a source and sink node that contains one or more cycles, we allow at most one traversal of each cycle.
2. We reject all paths that contain duplicate edges. This does not exclude paths with more than one cycle in which the cycles share a node, but does exclude paths with multiple cycles in which two or more cycles share an edge.

We eliminated paths containing more than one repetition of an internal cycle. For example, Figure 2 shows an example of a path with a cycle in the DNA damage network. (This is the *only* cycle in this network.) The shortest path from

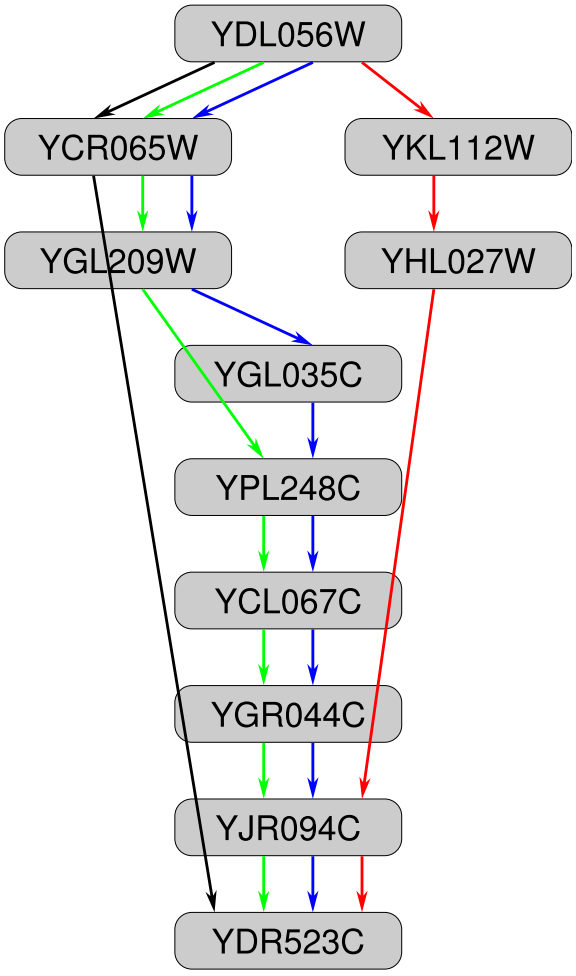


Figure 1. Multiple paths in the DNA damage network between source node YDL056W and sink node YDR523C: The paths contain 2, 4, 7 or 8 edges. Note that the two longest paths differ only in containing or skipping one intermediate node, *viz.,* YGL035C.

YJR060W to YPR158W is YJR060W → YBR049C → YOL004W → YGL073W → YPR158W. But, theoretically, an infinite number of other paths are possible by repeatedly traversing the 3-cycle YBR049C → YOL004W → YGL073W. In such cases, in enumerating the paths between two specific nodes we have retained only the paths containing no more than one traversal of the cycle (see caption to Figure 2.) (The 3-cycle is a special case of a feedback loop, and NOT an FFL motif.) We do not know whether the edges in these network graphs correspond to stimulatory or inhibitory interactions. However, if all the edges in a cycle were stimulatory, this would threaten runaway.

Figure 3 shows an example of a moderately complicated path in the cell-cycle network. This path contains two cycles.

Table 1 shows the basic statistics of different components of the five networks. The number of combinations – the product of the numbers of source and sink nodes – gives the number of pairs of nodes eligible to participate in source-to-sink paths. Each network contains one or more paths between only a small minority of these possibilities. For example, for the DNA damage network, there is a total of 37×643 (= number of source nodes×number of sink nodes) = 23791 possible combinations. For 95% of the combinations – 22511/23791 – there is no path within the network.

Despite having the fewest combinations, the cell-cycle network has the largest total number of paths and the largest percentage of possible paths.

For each of the networks, all source nodes are the start of at least one path to at least one sink node. For the cell-cycle, sporulation, and stress-response networks, all sink nodes are the termini of at least one path from at least one source node. For the DNA-damage and diauxic-shift networks, relaxing the constraint that reported paths have no repeated edges reveals paths that link source nodes to all sink nodes.

We next describe the usage of nodes, and the paths in the individual networks. We treat the DNA damage network in higher detail, as an example.

The DNA damage network

All of the 37 source nodes in the DNA damage network (those that have no parents) are the origins of paths to a sink node, or target gene. Table SM-1a (see supplementary material) lists these source nodes, and reports the number of paths between each of them and some sink node, and the range of lengths of these paths. In some cases, there are multiple paths between one source node and the same sink node (see Figure 1). There is a very great disparity of path initiations: the number of paths starting from a source node varies from 1 to a maximum of 308 (for YDL056W).

Table SM-1a reports the *total* number of paths starting at each source node, to any sink node. The number of paths between any particular source and sink node pair varies from 28 to 1.

Of the 643 sink nodes in the DNA damage network, 99 of them are the endpoints of paths from one or more source nodes. The maximum number of paths from all source nodes to one sink node is 20. For 58 of these 99 sink nodes, there are only 1 or 2 source-to-sink paths ending in them.

Table 2 shows the distribution of the numbers of paths between source and sink nodes, for sink nodes that are the endpoints of $\geqslant 4$ paths. (A complete table appears in the Supplementary

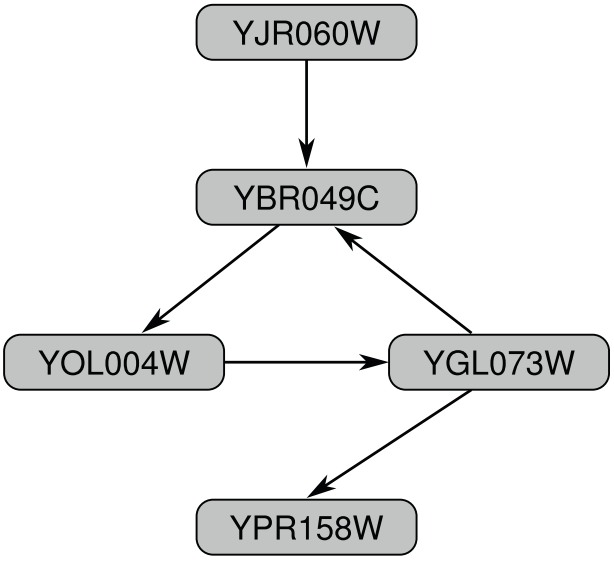


Figure 2. Path containing a cycle: If a directed graph contains a cycle, it is not possible to enumerate all paths between every pair of nodes. In the case illustrated here, possible paths in the DNA damage network from YJR060W to YPR158W include any number of repetitions of the cycle YBR049C → YOL004W → YGL073W. For this example we should report only two paths: YJR060W → YBR049C → YOL004W → YGL073 → YPR158W (*no* traversal of the cycle) and YJR060W → YBR049C → YOL004W → YGL073 → YBR049C → YOL004W → YGL073 → YPR158W (*one* traversal of the cycle).

Material, Table SM-2.) The sink nodes appear in decreasing order of the total number of paths from source nodes that end in them.

The cell-cycle network

All 16 source nodes in the cell-cycle network are the origins of paths to a sink node, or target gene. Table SM-1b (see supplementary material) lists these source nodes, and reports the number of paths between each of them and some sink node, and the range of lengths of these paths. Note the very great disparity of path initiations: the number of paths starting from a source node varies from 1, to a maximum of 131307 (for YDL056W).

The cell-cycle network contains a *much* greater number of source-to-sink paths than the other networks.

Table SM-1b reports the *total* number of paths starting at each source node, to any sink node. All of the 226 sink nodes in the cell-cycle network are endpoints of paths from one or more source nodes. That is, in this case, there are paths starting at all source nodes, and paths ending in each sink node; however, there are not paths joining every pair of source and sink nodes. The minimum number of paths from any source node to a particular sink node is 100: to sink nodes YMR198W, YLL040C, YML119W, and YNL262C. The maximum numbers of paths from all source nodes to a sink node is 24333 (to sink node YGL089C) and 23846 (to sink node YPL187W).

The sporulation network

All 29 source nodes in the sporulation network are the origins of paths to a sink node. Table SM-1c in the Supplementary material lists these source nodes, and reports the number of paths between each of them and some sink node, and the range of lengths of these paths. The number of paths starting from a source node varies from 1 to a maximum of 972 (for YDL056W).

Table SM-1c reports the *total* number of paths starting at each source node, to any sink node. The number of paths between any particular source and sink node pair varies from 56 to 1.

Of the 212 sink nodes in the sporulation network, 174 of them are the endpoints of paths from one or more source nodes. Of these, there is only 1 path to 8 of them. The maximum number of paths from all source nodes to one sink node is 163.

The diauxic-shift network

All the 35 source nodes in the sporulation network are the origins of paths to a sink node. Table SM-1d in the Supplementary Material lists these source nodes, and reports the number of paths between each of them and some sink node, and the range of lengths of these paths. The number of paths

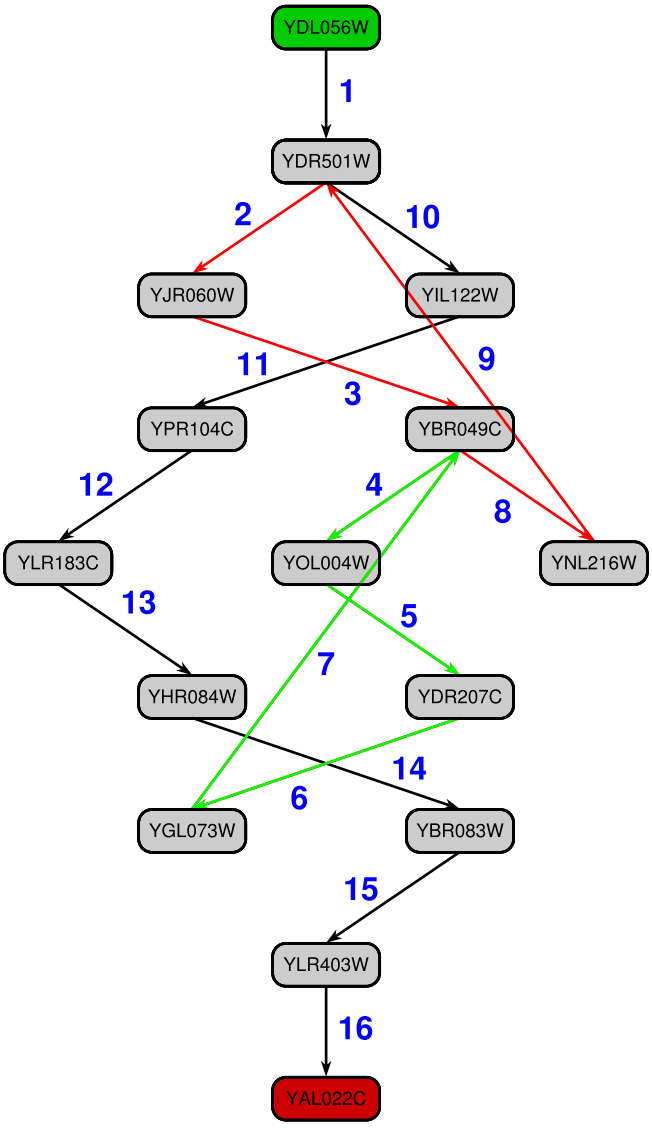


Figure 3. A typical path of moderate complexity, containing 16 edges, from the cell-cycle network, joining source node YDL056W (green) to sink node YAL022C (red): Edges are numbered consecutively in blue. The path contains two cycles: (YDR501W → YJR060W → YBR049C → YNL216W → YDR501C), shown by red arrows, and (YBR049C → YOL004W → YDR207C → YGL073W → YBR049C), shown by green arrows. Only one traversal of each cycle is shown. Obviously there is an infinite number of paths available by multiple traversals of either or both cycles. Note that the node YBR049C is common to both cycles. (In some cases cycles share edges; we have not retained such paths.)

Table 1 Numbers of nodes and edges, and source and sink nodes, in the five networks, and of possible, observed and absent paths

network	number of nodes	number of edges	number of source nodes	number of sink nodes	number of combinations	total number of source-to-sink paths	number of connected source-to-sink pairs	number of combinations for which no path found
cell cycle	296	550	16	226	3616	220319	1098	2518 (70%)
diauxic shift	783	1217	35	712	24920	2159	1329	23591 (95%)
DNA damage	715	1082	37	643	23791	1561	1280	22511 (95%)
sporulation	286	481	29	212	6148	3035	797	5351 (87%)
stress response	392	566	30	329	9870	1910	789	9081 (92%)

Table 2 Numbers of source-to-sink paths *to* each sink node, for sink nodes that are endpoints of $\geqslant 4$ paths from some source node. Source nodes appear in different columns; sink nodes in different rows. A full stop indicates that there is no path between a particular source node and a particular sink node. Total numbers of paths to each sink node appear at right of column

DNA damage

	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY BBBCDDDEEEEFGGGGHHIJJKKKKLLMMMNOOOPPPP LRRRLRRLRRLLLLLRRLLRLLLRRRRRRLLRRLLLL 00200340010001212100000002001110300 02 28995140863176320056334315248902537 83 13736089891316746160283436132988885 90 CWWWWCWCCWWCWWCWWCCWCWWWWWCWCCCCWWWCW	
YPL187W	....6..4.........14..6.7..1.....	29
YIL099W	....4..2........3..2..6.3111...1....	24
YJL148W	....3..1......1.2..4..2.2112.1.1..11.	23
YDR523C	....41.1..1.1......2..4.2..1...1.....	18
YHL022C	....3..1..1........3..3.2..1...1...1.	16
YER079W	....3..1........2..2..2.2..3.........	15
YHR014W	....3..1..1........3..3.2..1...1.....	15
YDR042C	.....1..........2..1..1..112.1.211...	14
YNL210W	....3..1..1........2..3.2..1...1.....	14
YOL126C	....5.............4..4..............	13
YIL162W	.1..2.1.........1..1..3....1.1.......	11
YLR367W	....3..1...........2..2.2..1.........	11
YIL119C	....1...........2...1.2..111...2.....	11
YNL338W	....3..1...1.......1..3.1...1........	11
YAL054C	....3..1...........5..2..............	11
YPL256C	....3..1...........2..2.2..1.........	11
YIL015W	....2..1...........2..2.1..1.......1.	10
YOR315W	..............3...1....1.2...2...1.	10
YDR077W	..............2.....1..111...2...1.	9
YIL111W	....1...........2...1.1..1.1...1....1	9
YMR017W	..............3.....2..111...1.....	9
YPR167C	.......3...........4.............2...	9
YNL336W	....2..1...........1..2.1...1........	8
YJR048W	1...2.........1.......2..1........1..	8
YDR146C	....2.............2..2....1..1......	8
YLR174W	....3.............3..2..............	8
YDR208W	....1...........2........111...1.....	7
YOR117W	.1..1...........1..3.......1.........	7
YFR030W	.......2...........4.............1...	7
YBR101C	....3..1...........1..2..............	7
YJR010W	.......2...........4.............1...	7
YLR460C	....1..1........1..1..1....1.........	6
YIL110W	....2.............2..2..............	6
YDR044W	..............2........2.1...1.....	6
YNL134C	....1...........1..1..1....1......1.	6
YDR345C	....2..............12........1.....	6
YJR095W	....2.............2..2..............	6
YHR007C	..............2........2.1...1.....	6
YLR081W	....2.............1..2........1.....	6
YCR107W	....2..1...........1..2..............	6
YBR266C	....2.............2..2..............	6
YGR088W	..............1..1..1..111..........	6
YCR039C	.......1...........11....3..........	6
YHR092C	....2..............12...........1..	6
YER007CA	..............2.......1.1.1.1.....	6
YKL148C	1...1.........1....2..1..............	6
YJR044C	....1..1...1.......1.......1.1.......	6
YNL081C	....2.............2..2..............	6
YML075C	..............2........2.1...1.....	6
YMR011W	....2..............13..............	6

	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY BBBCDDDEEEEFGGGGHHIJJKKKKLLMMMNOOOPPPP LRRRLRRLRRLLLLLRRLLRLLLRRRRRRLLRRLLLL 00200340010001212100000002001110300 02 28995140863176320056334315248902537 83 13736089891316746160283436132988885 90 CWWWWCWCCWWCWWCWWCCWCWWWWWCWCCCCWWWCW	
YBR126C	....2..............3..............	5
YEL040W	....................2...11.......1.	5
YMR105C	....2..............3..............	5
YDR157W	................1..3.......1........	5
YNL241C	.1..........1...1..........2........	5
YJR104C	....1........1..1..........1..1.....	5
YLR450W	..............2........1.1...1.....	5
YBL043W	.......1...........1..1...1....1.....	5
YGL055W	..............2........1.1...1.....	5
YOL151W	....1...........1.....1....1...1.....	5
YDL083C	......1...1.......1....1..........1.	5
YKL109W	....2..............3..............	5
YDR011W	...........1....1..1.......1.......1.	5
YFL014W	....1...........1..1..1....1........	5
YJR047C	..............2........1.1...1.....	5
YKL182W	....1..............2..........11.....	5
YEL070W	....2..............3..............	5
YBR021W	....2..............2............1.	5
YML056C	..............2........1.1...1.....	5
YNL277W	.......1...........2.......1.....1...	5
YGR059W	1...11........1.....................	4
YLR089C	....1..1...........1..1..............	4
YEL074W	.......1...........1..1.1...........	4
YLR272C	....1..1...........1..1..............	4
YBR050C	....2..............2..............	4
YAL040C	....1...........1....1....1.........	4
YDR516C	....2..............2..............	4
YML116W	..............1.....1....1...1.....	4
YDR010C	..............1..1.......1.......1.	4
YMR324C	....1..............1..1.1...........	4
YOR138C	.1.............1..........1.1.......	4
YGR149W	....1..1...........1..1..............	4
YHR094C	....2..............2..............	4
YNR018W	.1..1...........1..........1.........	4
YDR161W	....1..1...........1..1..............	4
YLL067C	....1..1...........1..1..............	4
YPL282C	....1..1...........1..1..............	4
YHR018C	.......2...........2..............	4
YGR234W	............1.1..........2.........	4
YOR202W	...1...1........1..........1.........	4
YIR029W	....2..............11..............	4
YOL119C	....1...........1.....1....1.........	4
YOR120W	....2..............2..............	4
YFL021W	.......1...........1...1...1.........	4
YPR013C	....1..............2.......1......	4
YIL018W	....1..1...........1....1............	4
YPR119W	..............1.2.......1.........	4
YJL212C	....1..1...........1.............1...	4
YGR055W	....2..1...........1.................	4
YKR093W	....1...........1.....1...1..........	4

starting from a source node varies from 1 to a maximum of 789 (for YGL096W).

Table SM-1d reports the *total* number of paths starting at each source node, to any sink node. The number of paths between any particular source and sink node pair varies from 21 to 1.

Of the 712 sink nodes in the diauxic-shift network, 525 of them are the endpoints of paths from one or more source nodes. Of these, there is only 1 path each to 234 of them. The maximum number of paths from all source nodes to one sink node is 45.

The stress-response network

All of the 30 source nodes in the stress-response network are origins of at least one path to a sink node. Table SM-1e in the Supplementary Material lists these source nodes, and reports the number of paths between each of them and some sink node, and the range of lengths of these paths. The number of paths starting from a source node varies from 1 to a maximum of 343 (for YMR043W).

Table SM-1e reports the *total* number of paths starting at each source node, to any sink node. The number of paths between any particular source and sink node pair varies from 18 to 1.

All of the 329 sink nodes in the stress-response network are the endpoints of paths from one or more source nodes. Of these, there is only 1 path to 143 of them. The maximum number of paths from all source nodes to one sink node is 89.

Comparison of paths among the networks

Table 1 reported the total number of source-to-sink paths in each network. Figure 4 shows the distribution of lengths (= number of edges) of the paths found. Note that each network has some paths containing only one edge; that is, an edge linking a source to a sink node directly.

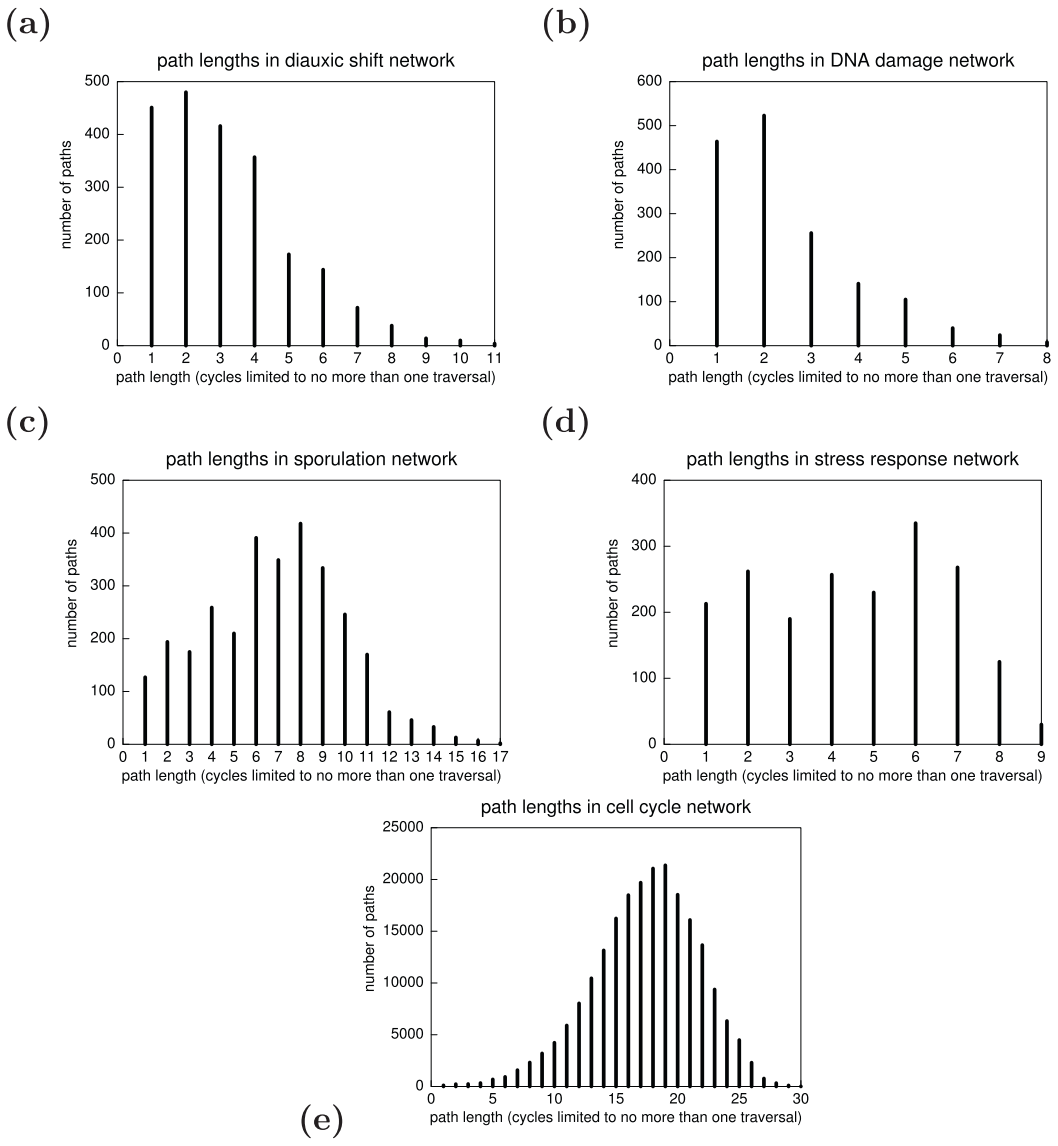


Figure 4. Distributions of source-to-sink path lengths in the different networks. (Path length = number of edges.) Clearly, the distributions of the number of paths as a function of path length are remarkably different. In particular, readers will be struck by the very different character of the cell-cycle network compared to the others.

It is striking that the cell-cycle network has properties very different from the others.

It is not apparent why the cell-cycle network network is so different from the others. It is *not* true that it has a larger number of nodes or edges than the other networks (see Table 1). Indeed, it is an outlier in having significantly *fewer* source nodes than the other networks.

A hypothesis to consider is that the diauxic shift, DNA damage, sporulation, and stress response networks govern behaviour that has one decision, controlling a single process directed to a state, or a single process together with its reverse, coupling *two* states. For instance, the diauxic shift network makes a transition between aerobic and anaerobic conditions. The cell-cycle network, on the other hand, has to integrate a linked succession of processes, and may therefore require a more complex network. We offer this suggestion together with the admission that we do not have what anyone would call hard evidence for it.

Multiplicities of source-sink pathways

To what extent are connected source and sink nodes controlled by unique, or only small numbers of paths? Multiple paths between a particular pair of source and sink nodes suggests a more complex control structure, in which different sets of intermediate nodes, along different paths, can differentially influence expression of the target gene. (But note, in Figure 1, that different

pathways can show significant overlap, sharing intermediate nodes.)

Different networks have different distributions of multiplicities of source-sink paths. Table 1 shows, for each network, the total number of source-sink pairs connected by paths.

Among networks, and within each network, there is very great disparity in the numbers of source-sink paths. For four of the five networks – except for the cell-cycle network – there are relatively few paths between many connected pairs of source and sink nodes.

- For the DNA-damage network, only 6 of the 1280 connected source-sink pairs have $\geqslant 5$ paths between them (maximum = 7); 1077 source-sink pairs are connected by only 1 path (152 pairs are connected by 2 paths).
- For the diauxic-shift network, only 13 of the 1329 connected source-sink pairs have $\geqslant 10$ paths between them (maximum = 21); 965 source-sink pairs are connected by only 1 path (206 pairs are connected by 2 paths.)
- For the stress-response network, only 14 of the 2006 connected source-sink pairs have $\geqslant 10$ paths between them (maximum = 18); 476 source-sink pairs are connected by only 1 path (106 pairs are connected by 2 paths).
- For the sporulation network, 33 of the 797 connected source-sink pairs have $\geqslant 20$ paths between them (maximum = 56); 415 source-sink pairs are connected by only 1 path (94 pairs are connected by 2 paths).
- The cell-cycle network shows different behaviour: 43 of the 789 connected source-sink pairs have $\geqslant 1000$ paths between them (maximum = 14522, followed by

Table 3 Cyclic subpaths in the different networks

DNA damage
(YBR049C → YOL004W → YGL073W → YBR049C)

diauxic shift
(YOL004W → YGL073W → YBR049C → YOL004W)
(YBR049C → YOL004W → YGL073W → YBR049C)

sporulation
(YGL073W → YBR049C → YOL004W → YGL073W)
(YBR049C → YOL004W → YDR207C → YGL073W → YBR049C)
(YDR207C → YGL073W → YBR049C → YOL004W → YDR207C)
(YBR049C → YNL216W → YGL073W → YBR049C)
(YGL073W → YBR049C → YNL216W → YGL073W)
(YGL073W → YBR049C → YOL004W → YDR207C → YGL073W)
(YBR049C → YOL004W → YGL073W → YBR049C)
(1 YPR065W → YDR259C → YPR065W 1)
(1 YDR259C → YOR028C → YDR259C 1)
(1 YDR259C → YPR065W → YDR259C 1)
(1 YBR049C → YOL004W → YGL073W → YBR049C 1)

stress response
(1 YPR065W → YDR259C → YOR028C → YPR065W 1)
(1 YDR259C → YOR028C → (2 YPR065W → YDR259C 1) → YPR065W 2)
(1 YDR259C → YOR028C → YPR065W → YDR259C 1)
(1 YOL004W → YGL073W → YBR049C → YOL004W 1)
(1 YPR065W → (2 YDR259C → YOR028C → YDR259C 2) → YPR065W 1)
The third cycle in this list YDR259C → YOR028C → YPR065W → YDR259C is contained in the second cycle in the list.
In different paths, this cycle appears in both on its own and embedded in a longer cycle.

cell cycle
The shortest cycles in the cell-cycle network.
(1 YBR049C → YNL216W → YGL073W → YBR049C 1)
(1 YBR049C → YOL004W → YGL073W → YBR049C 1)
(1 YDR501W → YIL122W → YPR104C → YDR501W 1)
(1 YGL073W → YBR049C → YNL216W → YGL073W 1)
(1 YGL073W → YBR049C → YOL004W → YGL073W 1)
(1 YIL122W → YPR104C → YDR501W → YIL122W 1)
(1 YPR104C → YDR501W → YIL122W → YPR104C 1)
(1 YBR049C → YNL216W → YDR501W → YJR060W → YBR049C 1)
(1 YBR049C → YOL004W → YDR207C → YGL073W → YBR049C 1)
(1 YDR207C → YGL073W → YBR049C → YOL004W → YDR207C 1)
(1 YDR501W → YJR060W → YBR049C → YNL216W → YDR501W 1)
(1 YKL112W → YPR104C → YLR183C → YLR182W → YKL112W 1)
(1 YPR104C → YLR183C → YLR182W → YKL112W → YPR104C 1)
(1 YBR049C → YNL216W → (2 YGL073W → YBR049C 1) → YOL004W → YGL073W 2)
(1 YBR049C → YOL004W → (2 YGL073W → YBR049C 1) → YNL216W → YGL073W 2)

14226, all with source node YDL056W); 328 have $\geqslant$ 100 paths between them; only 94 source-sink pairs are connected by only 1 path (only 4 pairs are connected by 2 paths). YDL056W is the source node for 131307 source-sink paths in the cell-cycle network; this is over half of all the source-sink paths in this network. Of the 14522 paths between source node YDL056W and sink node YGL089C, 7282 (almost exactly half) have lengths $\geqslant$ 20.

Cycles in the paths in the different networks

Table 3 lists all the cycles found in all networks other than the cell-cycle network, for which only the shortest cycles are shown. The cell-cycle network has 298 cycles; the longest contains 18 edges. (A complete table appears in the supplementary material, Table SM-3.)

Multiple overlapping cycles are parenthetically delimited using corresponding numeric counters, for example (1 . . . (2 . . . 2) . . .1).

There is a total of 305 unique cycles in the five networks. Of these, only one appears in all five:

(YBR049C → YOL004W → YGL073W → YBR049C)

Seven cycles appear in two of the networks, all but one in the cell-cycle and sporulation networks:

(YGL073W → YBR049C → YOL004W → YGL073W)	cell cycle	sporulation
(YGL073W → YBR049C → YOL004W → YDR207C → YGL073W)	cell cycle	sporulation
(YGL073W → YBR049C → YNL216W → YGL073W)	cell cycle	sporulation
(YDR207C → YGL073W → YBR049C → YOL004W → YDR207C)	cell cycle	sporulation
(YBR049C → YOL004W → YDR207C → YGL073W → YBR049C)	cell cycle	sporulation
(YBR049C → YNL216W → YGL073W → YBR049C)	cell cycle	sporulation
(YOL004W → YGL073W → YBR049C → YOL004W)	diauxic shift	stress response

The other 297 cycles are unique to a particular network.

All the cycles that appear in more than one network contain the edge.

YGL073W → YBR049C; five of the eight contain the edge YBR049C → YOL004W.

Do the networks contain hubs?

In addition to the general question of how the networks overlap, we asked whether there are 'hubs'; that is, nodes that participate in an unusually large number of edges, and whether these are shared by multiple networks. To address this question we counted the usage, in the source-to-sink paths, of each node in each network, including internal nodes that are neither

Table 4 Top node usage in different networks

cell cycle		diauxic shift		DNA damage		sporulation		stress response	
411198	YDR501W	789	YGL096W	342	*YCR065W*	1807	**YBR049C**	1924	YDR259C
338135	YPR104C	602	**YBR049C**	308	**YDL056W**	1203	**YGL073W**	882	YPR065W
326209	**YBR049C**	497	YDR146C	260	**YJR060W**	1123	YDR207C	752	YOR028C
198686	YLR183C	417	**YJR060W**	231	**YKL043W**	972	**YDL056W**	436	YPL177C
197044	YER111C	406	**YOL004W**	191	**YBR049C**	960	*YCL067C*	403	YIL122W
192114	YLR182W	364	**YDL056W**	186	YNL103W	757	YNL216W	343	YMR043W
188742	YML027W	358	**YGL073W**	178	YGL209W	740	YLR182W	338	YLR131C
186492	**YJR060W**	346	YGL209W	148	YIR018W	737	**YJR060W**	323	YML007W
181027	YNL216W	272	YLR013W	136	YMR043W	732	YML027W	316	**YKL043W**
164801	YIL122W	245	YGR044C	130	YEL009C	728	YJR094C	275	**YBR049C**
145006	**YGL073W**	218	YML027W	110	YHR206W	722	YKL112W	272	YGL096W
141920	YKL062W	175	YGL035C	104	YML007W	705	YDR146C	250	**YDL056W**
131307	**YDL056W**	171	YMR043W	103	**YOL004W**	684	YPR104C	235	YDR043C
117910	**YOL004W**	166	YNL103W	102	YKL112W	672	*YCR065W*	223	YMR021C
100900	YDR146C	160	*YCL067C*	98	YLR131C	654	**YKL043W**	218	YDR501W
92650	**YKL043W**	156	YIR018W	94	YGL035C	648	YEL009C	201	**YJR060W**
79702	YDR207C	147	YKL109W	88	YPL248C	637	YGR044C	184	**YOL004W**
58950	*YCL067C*	112	YEL009C	85	YPR065W	616	**YOL004W**	154	**YGL073W**
58522	YLR013W	108	YKL112W	84	**YGL073W**	603	YGL209W	136	*YCR065W*
57638	*YCR065W*	99	**YKL043W**	84	*YCL067C*	585	YNL103W	96	YLR256W

source nor sink nodes. Table 4 shows the top 20 nodes in each network. (For this calculation, we counted parent and child nodes in edges equivalently.) Source nodes appear in blue. The 6 nodes that are among the top 20 nodes in all 5 networks – YBR049C, YJR060W, YGL073W, YOL004W, YDL056W, and YKL043W – appear in boldface. Note that YDL056W is a source node in all five networks. YJR060W is a source node in the DNA damage and sporulation networks. YKL043W is a source node in the stress-response network but not in the others. Two nodes that are among the two 20 nodes in only four of the 5 networks – YCL067C and YCR065W – appear in italics. Seven nodes appear among the two 20 nodes in only three of the five networks; fourteen nodes appear among the top 20 nodes in only two of the five networks; and thirteen nodes appear among the top 20 nodes in only one of the five networks. Most of the nodes that appear in among the top 20 nodes in only one of the networks do appear within other networks, but not within the top 20 nodes in more than one of them.

Materials and Methods

We used the transcription regulatory networks of *Saccharomyces cerevisiae* under various physiological conditions published by Luscombe and coworkers: http://networks.gersteinlab.org/regulation/dynamics/index2.html.

We wrote programs to identify, in each network, the source and sink nodes, and to enumerate, for each network, possible paths between each of the source nodes and each of the sink nodes (see http://www.bx.psu.edu/aml2/generatepaths.prl).

Discussion

We ask, to what extent do the five networks overlap? Are they largely disjoint and independent; that is, containing separate sets of edges, nodes and paths? Or do these sets overlap: To what extent do the networks construct themselves from the same set of materials – nodes and edges? To what extent do they assemble common materials in the same ways – into common paths?

Altogether the five networks contain 2479 edges. Only 23 edges are common to all five networks. 43 edges appear in four of the five networks, but not in all five. 259 edges appear in three of the five networks, but not in more. 678 edges appear in two of the five networks, but not in more. 1476 edges (60%) appear in only one of the five networks. The cell-cycle network contains a total of 550 edges, of which 278 appear in at least one of the other networks. Only 272 edges are unique to the cell-cycle network (very close to half the total number of nodes in the cell-cycle network).

Sharing of all nodes and edges among networks

Table 5 shows the sharing of nodes and edges between pairs of networks. The number in parentheses are the total number of nodes or edges in the individual networks. This includes *all* nodes, including those internal to paths, not only the source and sink nodes. The results are presented as follows: For instance (upper left numerical entry in Table): The DNA damage and sporulation networks individually involve 715 and 286 nodes, respectively, participating in 1082 and 481 edges, respectively. These two networks share 123 nodes and 174 edges. These results imply the following statements about this pair of networks:

- In all five networks taken together, there is a total of 1475 nodes and 2479 edges.
- The DNA damage network uses 715 nodes (48% of the total), and contains 1083 edges (43% of the total).
- The sporulation network uses 286 nodes (19% of the total), and 481 edges (19% of the total).
- The DNA damage and sporulation networks share 123 nodes (8% of the total, but 17% of all the nodes appearing in the DNA damage network and 43% of all the nodes appearing in the sporulation network.)
- The DNA damage and sporulation networks share 174 edges (7% of the total, but 16% of all the edges appearing in the DNA damage network and 36% of all the nodes appearing in the sporulation network.)

Pairs of networks are certainly not disjoint; they do share nodes and edges, but to a lesser extent

Table 5 Sharing of nodes and edges between different networks

	sporulation	stress response	diauxic shift	cell cycle
DNA damage	123(715/286) 174(1082/481)	202(715/392) 258(1082/566)	337(715/783) 476(1082/1217)	122(715/296) 166(1082/550)
sporulation		56(286/392) 54(481/566)	121(286/783) 154(481/1217)	85(286/296) 127(481/550)
stress response			268(392/783) 353(566/1217)	49(392/296) 45(566/550)
diauxic shift				104(783/296) 136(1217/550)

than the sharing of source and sink nodes. The most extensive sharing is seen in the DNA-damage and diauxic-shift networks, and the stress-response and diauxic-shift networks. The DNA-damage and diauxic-shift networks share 47% (DNA damage) and 43% (diauxic shift) of their nodes, and 44% (DNA damage) and 39% (diauxic shift) of their edges. The stress-response and diauxic-shift networks share 68% (stress response) and 34% (diauxic shift) of their nodes, and 62% (stress response) and 29% (diauxic shift) of their edges.

In summary, the networks do share nodes and edges, in some but not all cases quite robustly.

Sharing of source and sink nodes

In the five networks combined, there is a total of 58 unique source nodes. With the exception of the cell-cycle network, each network uses no less than half of them: between 50% and 64%.

In the five networks combined, there is a total of 1378 unique sink nodes. Each network uses between 15% and 52% of them.

Table 6 shows the sharing of source and sink nodes between pairs of networks. For instance:

- the cell-cycle network contains 16 source nodes and 226 sink nodes (cell [1, 1] of the matrix in Table 6).
- the DNA-damage network contains 37 source nodes and 643 sink nodes (cell [2, 2] of the matrix in Table 6).
- the DNA-damage network and cell-cycle networks share 15 source nodes and 66 sink nodes (cell [1, 2] of the matrix in Table 6).

In almost all cases, pairs of networks share no less than approximately half their source nodes. That is, the number of shared source nodes is approximately half the number of source nodes used by each of the two individual networks (not half the total number of source nodes).

The extent of sharing of sink nodes is, in general, smaller and more variable. The smallest extent of sharing is between the sporulation and stress-response networks: 11 shared sink nodes out of 212 (sporulation) and 329 (stress-response), corresponding to 5% and 3%. The largest extent of sharing is between the DNA-damage and diauxic-shift networks: 269 shared sink nodes out of 643 (DNA damage) and 712 (diauxic shift), corresponding to 42% and 38%.

In summary, the networks show to some extent common source and sink nodes, in many cases approximately half the source and sink nodes in pairs of networks are shared.

Sharing of paths among different networks

The sets of paths found in the different networks overlap to a limited extent: 979 of the total of 227839 unique source-to-sink paths appear in more than one network. No path appears in all five networks. This would not be possible, as, although five nodes appear as sources in all five networks – YDL056W, YGL237C, YMR021C, YMR043W, and YPL089C – no node appears as a sink in all five networks. Nine nodes appear as sinks in four of the five networks: YGL116W, YGR088W, YNR001C, YML091C, YAR007C, YNL312W, YKL142W, YHL028W, and YER070W.

Eleven single-edge source-to-sink paths appear in four of the five networks. Two two-edge source-to-sink paths, which share their first edge, appear in four of the five networks:

YDL056W → YKL112W → YNL312W
(in all but stress-response network)

YDL056W → YKL112W → YHL028W
(in all but stress-response network)

There are 140 paths that appear in three of the five networks. All have length 1, 2, or 3, except for:

YMR043W → YLR131C → YML007W → YPR065W → YJL148W	(DNA damage, diauxic shift, stress response)
YDL056W →YML027W →YKL043W →YKL109W →YML091C	(cell cycle, diauxic shift, sporulation)
YDL056W → YCR065W → YGL209W → YGL035C → YDR516C	(DNA damage, sporulation, stress response)

The node YDL056W appears as the source node in four of these five paths. Although identified in the *Saccharomyces* Genome Database as encoding a transcription factor involved in regulation of cell cycle progression from G1 to S phase, it is active in other networks also.

Table 6 Sharing of source/ sink nodes between different networks.

	cell cycle	DNA damage	sporulation	stress response	diauxic shift
cell cycle	**16/226**	15/66	14/32	7/12	13/52
DNA damage		**37/643**	20/61	14/157	27/269
sporulation			**29/212**	14/11	18/59
stress response				**30/329**	17/219
diauxic shift					**35/712**

Table 7 Number of paths shared between pairs of networks

	sporulation	stress response	diauxic shift	cell cycle
DNA damage	136 (3.0%)	208 (5.8%)	311 (8.4%)	88 (0.04%)
sporulation		14 (0.3%)	76 (1.5%)	92 (0.04%)
stress response			301 (7.2%)	14 (0.006%)
diauxic shift				84 (0.04%)

Table 7 shows the numbers of source-to-sink paths shared by pairs of networks. It reports the ratio of the number of shared paths to the total number of paths appearing in either network, expressed as a percentage. Many of these are short but there are 5 paths of length 9, all shared by the sporulation and cell-cycle networks, and all starting at node YDL056W.

Although the cell-cycle network is itself very rich in paths, it does not share them very generously with other networks.

Sharing of subpaths containing two consecutive edges

To explore how the different networks assemble common nodes and edges, we extracted from all paths a set of consecutive triples of nodes; that is, subpaths containing two consecutive edges. Thus, from a path A → B → C → D → E we should extract A → B → C, B → C → D, and C → D → E.

There is a total of 3062 unique triples in all five networks combined. The distribution is as follows:

network	number of unique triples
cell cycle	957
diauxic shift	1156
DNA damage	904
sporulation	629
stress response	572

Approximately three-quarters of the total of 3062 are unique to one of the networks:

Number of networks	Number of common triples
5	9
4	34
3	209
2	600
1	2210

The nine triples common to all five networks are:

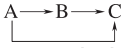
YOL004W	→	YGL073W	→	YOR344C
YOL004W	→	YGL073W	→	YBR049C
YLR013W	→	YCR097W	→	YGR044C
YLR013W	→	YCL067C	→	YGR044C
YJR060W	→	YBR049C	→	YOL004W
YGL209W	→	YGL035C	→	YKL109W
YDL056W	→	YKL112W	→	YIR023W
YBR049C	→	YOL004W	→	YGR044C
YBR049C	→	YOL004W	→	YGL073W

YDL056W is a source node in all five networks. YJR060W is a source node in the DNA-damage and sporulation networks. YLR013W is a source node in the DNA-damage network, but not in the others. YGL209W is a source node in the diauxic-shift network, but not in the others.

Table 8 shows the extent of pairwise sharing of two-consecutive-edge subpaths.

These results suggest that, even at the elementary level of 3-node subpaths, in most but not all cases the different networks assemble common nodes and edges in individual ways.

Sharing of Feed-Forward Loops

Some of the 4218 subpaths containing two consecutive edges (3062 unique ones) in the networks are parts of a motif: the feed-forward loop (FFL). Recall that an FFL has the form  in which there are paths of both length 1 and length 2 between nodes A and C. Altogether the networks contain 346 FFLs; with the following distribution:

Network	**Number of FFLs**
cell cycle	103
diauxic shift	64
DNA damage	70
sporulation	67
stress response	42
Total	**346**

Table 8 Sharing of two consecutive-edge subpaths between different networks

	cellcycle	diauxic shift	DNA damage	sporulation	stress response
cellcycle	**957**	142	94	147	34
diauxicshift		**1156**	322	158	227
DNAdamage			**904**	147	210
sporulation				**629**	40
stressresponse					**572**

Table 9 Sharing of FFLs between different networks

	cell cycle	diauxic shift	DNA damage	sporulation	stress response
cell cycle	**103**	9	16	17	1
diauxic shift		**64**	30	11	14
DNA damage			**70**	23	13
sporulation				**67**	5
stress response					**42**

Among the combined total of 346 there are 243 unique FFLs. The FFL 

is the only one common to all 5 networks. YGL209W is a source node in the diauxic-shift network only.

Five FFLs are common to four networks; four of them begin with the edge YGL209W → YGL035C, as does the FFL common to all five. Three of the five are absent from the cell-cycle network; two are absent from the stress-response network. Fifteen FFLs are shared by three networks, 54 are shared by 2 networks, and 168 appear in only one of the networks.

Table 9 shows the sharing of FFLs between pairs of networks. (The diagonal elements – *e.g.,* cell cycle/cell cycle 103 – show total numbers of FFLs in each network.) Only for the diauxic-shift and DNA-damage networks does the sharing of FFLs approach half; for other pairs it is considerably less.

Conclusions

1. The cell-cycle network is not an outlier in terms of total number of nodes and edges, and number of sink nodes. It has slightly fewer source nodes than the others. However, the cell-cycle network has almost an order of magnitude more source-to-sink paths than the other networks.
2. The networks are not disjoint and independent, but show substantial although not anywhere near complete overlap in their usage of nodes and edges:
   - The usage and sharing of nodes and edges is much higher than the usage and sharing of paths.
   - The networks contain paths between rather small fractions of potential source-sink pairs. The cell-cycle network has the most (30%), the sporulation network has 13%, but the diauxic-shift, DNA-damage and stress-response networks have 8% or fewer.
   - The networks show very high variability in the numbers of paths linking different source to sink nodes. The cell-cycle network also shows far greater multiplicity of paths between particular source and sink nodes than the other networks (this is *not* a necessary consequence of the high path/node ratio in the cell-cycle network, because, in the cell-cycle network there are NO paths between 70% of the possible source-sink pairs).
   - Although the networks do contain hubs, few nodes act as hubs in multiple networks.
3. The picture that emerges is that the networks make use of overlapping sets of materials – nodes and edges – but assemble them independently into different sets of paths.

CRediT authorship contribution statement

Arthur M. Lesk: Conceptualisation, Methodology, Software, Validation, Visualisation, Project administration. **Arun S. Konagurthu:** Software, Validation, Writing – review & editing.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

.

Appendix A. Supplementary Material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jmb.2021.167181.

Received 19 June 2021;
Accepted 27 July 2021;
Available online 30 July 2021

Keywords:
regulatory networks;
pathway analysis and comparison;
systems biology

References

1. Lesk, A.M., Konagurthu, A.S., (2020). Neighbourhoods in the yeast regulatory network in different physiological states. *Bioinformatics*, **37**, 551–558.
2. Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M., (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

3. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
4. Konagurthu, A.S., Lesk, A.M., (2008). Single and multiple input modules in regulatory networks Proteins: Struct., Funct. *Bioinform.*, **73**, 320–324.
5. Konagurthu, A.S., Lesk, A.M., (2008). On the origin of distribution patterns of motifs in biological networks. *BMC Syst. Biol.*, **2**, 73.
6. Yu, H., Gerstein, M., (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. USA*, **103**, 14724.
7. Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M., Aravind, L., (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.