PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Single and multiple input modules in regulatory networks

Arun S. Konagurthu* and Arthur M. Lesk*

The Huck Institute for Genomics, Proteomics, and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania

ABSTRACT

Interactions between transcription factors and target genes form regulatory networks that control target gene expression. Regulatory networks contain canonical motifs, including the feed forward loop (FFL), single input module (SIM), and multiple input module (MIM) (Fig. 1). A challenge for network analysis is to identify and enumerate the motifs, required to illuminate their biological significance. Although there is consensus about the definition of the FFL, published definitions of the SIM and MIM are unclear and often used inconsistently. Here, we provide, for the first time, a complete and consistent definition of SIM and MIM, and algorithms for enumerating SIMs and MIMs in any network. From the algorithmic point of view, enumeration of SIMs and MIMs is substantially harder than enumerating FFLs. We compare the distributions of motifs in the Yeast regulatory network under different physiological conditions, reported earlier by the landmark paper of Luscombe et al. ***(Nature 2004, 431: 308–312). Our reanalysis shows major differences in the number of motifs compared with the results of those authors, requiring significant revision of some of their conclusions.***

Proteins 2008; 73:320–324.

Key words: ***motifs; transcription regulatory network; SIM; MIM; FFL.***

INTRODUCTION

Metabolic and regulatory networks are directed graphs. Motifs are subgraph patterns that occur at higher frequencies than expected in random networks with similar connectivity parameters. By a subgraph, we mean a subset of nodes, retaining all of the original arcs (*directed edges*) between them.

Pioneering work of Alon and coworkers identified several motifs in regulatory networks and ascribed informational processing roles to them.[2–4] A landmark study by Luscombe *et al.* reported that the yeast regulatory network is "rewired" in different physiological states, with changes in the distribution of motifs. This gave important clues to the nature and mechanism of the response to changing conditions.[1]

Many authors take as a canonical set of network motifs: the single-input motif (SIM), the feed-forward loop (FFL), and the multiple-input motif (MIM) (see Fig. 1).

There is consensus that a FFL is a set of three nodes with both direct and indirect paths from one of the nodes to a second.

However, there surprisingly appear to be no precise definitions of the SIM and MIM. Authors generally offer no more than an anecdotal definition by reproducing pictures from Alon's work [equivalent to Fig. 1(a,c)], and do not always even interpret them consistently.

A common feature of definitions of SIMs is that a SIM contains arcs from one primary node to at least two secondary nodes, with no additional input from any other node to the secondary nodes. However, in enumerating SIMs, some authors impose a further constraint that the secondary nodes do not have any further outgoing edges.[1] The only motivation for this additional constraint appears to be to simplify the enumeration problem.

The confusion intensifies when it comes to MIMs. Alon and coworkers defined the graph of Figure 1(c)—containing exactly two input nodes and two output nodes—as a *bifan*. The MIM emerged in the literature as an unspecified generalization of the bifan.[5] Alon and coworkers also defined a related motif, the dense overlapping regulon (DOR), as a two-layer subnetwork, with many but not necessarily complete connections between the layers but none within either layer, although the threshold density of connections is not specified.[2,4]

Perhaps, the elusiveness of the definitions is related to the fact that it is harder to devise algorithms for counting SIMs and MIMs than for counting FFLs. However, it is essential to have unambiguous definitions, if only to be able to compare and reproduce the work by the various investigators active in the field. We shall

*Correspondence to: Arun S. Konagurthu or Arthur M. Lesk, The Huck Institute for Genomics, Proteomics, and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802. E-mail: arun@bx.psu.edu or aml25@psu.edu

Received 21 December 2007; Revised 5 February 2008; Accepted 25 February 2008
Published online 23 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22053

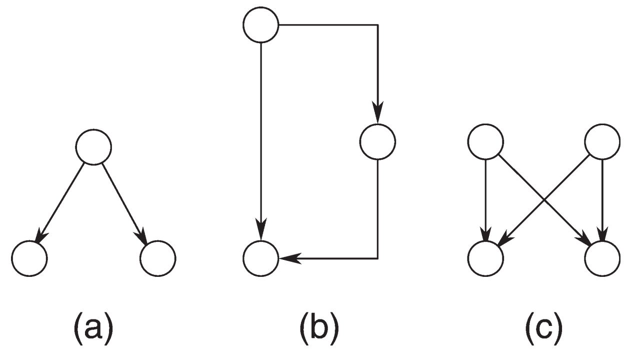

Figure 1
*Canonical motifs in regulatory networks, (**a**) single-input motif (SIM), (**b**) feed-forward loop (FFL), and (**c**) Multiple-input motif (MIM).*

see that, in the yeast regulatory network at least, the biological message depends on consistent application of the definitions of the motifs used in analyzing the data.

Here we give formal definitions of SIM and MIM, consistent with the general usage of the terms in the literature, and present effective algorithms for counting SIMs and MIMs. These are not the only possible definitions. However, we strongly emphasize that only by consistently applying precise definitions, can we learn whether they can be improved, based on the criterion of maximizing the biological insight that network analysis provides.

METHODS

Definitions

Single and multiple input modules (SIM and MIM) in a directed graph are *maximal* subgraphs comprising two nonempty disjoint sets (layers): $\mathcal{P}$ and $\mathcal{C}$ (standing for Parent and Child).

A SIM requires that $\mathcal{P}$ contain only one node and $\mathcal{C}$ contain at least two nodes, such that the full graph contains an arc from the parent node to every $c_i \in \mathcal{C}$. We also require the *indegree* — number of incoming edges — of every c_i to be strictly equal to one, within the full network, not just within the subgraph. By this definition of a SIM, no edges can exist between any $c_i, c_j \in \mathcal{C}$. It follows that $\mathcal{P}$ is the only parent of all nodes in set $\mathcal{C}$.

A MIM requires that both $\mathcal{P}$ and $\mathcal{C}$ must contain ≥ 2 nodes, and that there is an arc from every $p_i \in \mathcal{P}$ to every $c_i \in \mathcal{C}$, no edge between any $p_i, p_j \in \mathcal{P}$, and no edge between any $c_i, c_j \in \mathcal{C}$.

These definitions provide a set of fundamental network motifs that are, in a sense, "orthogonal:" No subgraph can be more than one of the SIM, FFL, and MIM set of motifs.

Authors do not consistently impose the criterion of maximality in enumerating SIMs and MIMs. However, a maximal MIM with P parents and C children contains $[2^P - (P + 1)] \times [2^C - (C + 1)] - 1$ easily enumerable non-maximal "subMIMs". Similarly, a maximal SIM with one parent and C children contains $2^C - C - 2$ non-maximal "subSIMs."

Our definition of the SIM excludes the counting of subgraphs of MIMs as SIMs. Without the indegree constraint on a SIM, every MIM containing P parents and C children would contain P subgraphs that form either a SIM or a subset of a SIM. Therefore the indegree constraint is essential for the independence of the motifs.

Of the motifs, MIMs provide the highest channel capacity from stimulus to response. If all the links from parents to children are stimulatory, MIMs allow robust mobilization of many target genes by many different individual stimuli: distributed control. Alternatively, if the links from some parents are uniformly stimulatory and those from others are uniformly repressive, then a MIM can function as a "gang switch", turning on or off a large number of target genes in parallel. Many more complex modes of control are of course also possible.

Algorithm to find all SIMs in a directed network

Every arc in the directed graph joins a source node, called the parent, to a target node, the child. Any node may be both the source of one or more arcs, and the target of one or more arcs; that is, any node may be the parent in many edges and the child in many edges. (We do not consider edges between a node and itself). If there is an arc from node N_1 to node N_2, N_1 is a parent of N_2 and N_2 a child of N_1. Let $P(N_i)$, the set of all parents of N_i, and $C(N_i)$, the set of all children of N_i.

Buildup of maximal subset $\mathcal{C}$ in a SIM

Each node N_i is a potential parent node in a SIM. For each node N_i in the full graph find the set $C(N_i) \equiv C_i$ of all its children (=the set of nodes that are targets of an arc from N_i). Discarding singletons, shrink the set C_i by eliminating nodes which have incoming edges from nodes outside $N_i \cup C_i$. The reduced C_i, provided it contains at least two elements, forms the candidate set from which maximal independent sets of nodes—which is, set of nodes without links between them—are extracted.

Finding maximal independent subsets of reduced C_i

We make use of the well-known Bron and Kerbosch[6] algorithm for finding cliques in graphs. A clique is a fully connected subgraph, a subgraph with an edge between every two nodes.

Consider the (undirected) complement of the graph induced by a reduced C_i. This is created by deleting all edges originally present, and introducing an (undirected) edge between any two nodes originally unconnected. Then the condition that a subgraph of the reduced C_i set has no edges becomes that the corresponding subgraph of the inverted graph is completely connected.

The Bron and Kerbosch method[6] is used to solve the problem of finding the maximal independent subsets of nodes in the secondary layer of a SIM, that contain no interchild edges. All cliques (of size ≥ 2 nodes) in the complement graph gives us the maximal sets of independent child nodes of a SIM for every parent node N_i.

Algorithm to find all MIMs in a directed network

Using the same notation as in "Algorithm to find all SIMs in a directed network" section, our algorithm first identifies maximal candidate $\mathcal{P}$ and $\mathcal{C}$ subsets that contain arcs linking every element of the first subset to every element of the second. From these candidates, we extract maximal subsets, each free of links within $\mathcal{P}$ and $\mathcal{C}$.

Buildup of maximal subsets $\mathcal{P}$ and $\mathcal{C}$

For each node N_i find the set C_i of all its children (= the set of nodes that are targets of an arc from N_i). For each pair $c_1, c_2 \in C_i$, find the maximal subset of nodes that are parents of c_1 and c_2; that is, the intersection of the parents of c_1 and the parents of c_2: $P(c_1) \cap P(c_2)$.(If this intersection set has at least two elements, it, and all of its subsets containing ≥ 2 nodes, are candidate parent sets of a MIM, but we have to check that there is no edge between c_1 and c_2, nor between any pair of the mutual parents of c_1 and c_2.) We take *pairs* of elements from each C_i to ensure that every MIM have at least two child nodes. Similarly, we require that the intersected parent sets $P(c_1) \cap P(c_2)$ have ≥ 2 elements.

By considering all pairs of elements from the child sets derived from every node, we build up a list of child and corresponding parent sets such that all parents have arcs to all children. Initially, c_1,c_2 and $P(c_1) \cap P(c_2)$ form a child and corresponding parent set (provided that $P(c_1) \cap P(c_2)$ contains $\geq$ 2 nodes). If another pair c_3, c_4 gives us the same intersection of parents —$P(c_1) \cap P(c_2) = P(c_3) \cap P(c_4)$— we merge the child sets, updating our list to contain: c_1, c_2, c_3, c_4, and $P(c_1) \cap P(c_2) (= P(c_3) \cap P(c_4))$, We build up maximal child and corresponding parent sets. The process is repeated for all pairs in the parent sets that share the same children.

This step has built up candidate MIMs, that are complete as far as parent-child arcs are concerned. The next step is as follows.

Finding subsets of candidate MIMs free of parent–parent and child–child edges

Consider the parent set in a candidate MIM. "Invert" this graph by deleting all edges originally present and introducing an edge between any two nodes originally unconnected. Then, the condition that a subgraph of the parent set has no edges becomes that the corresponding subgraph of the inverted graph is completely connected. We invert the child set also. We do not change any edges linking nodes in the parent set to nodes in the child set; we know by construction that there is an edge between every parent and every child node.

Then, the problem of finding the maximal subsets of parent and child graphs, that contain no interparent edges and no interchild edges but all parent–child edges, is equivalent to finding all cliques in every candidate MIM derived in the first step, after inverting all interparent and interchild edges. This is the problem solved by Bron and Kerbosch.[6]

We note that finding cliques is NP-hard.[6] However, we observe that in real regulatory networks, the sizes are SIMs and MIMs are small enough to use the Bron and Kerbosch algorithm practically.

Note on enumeration of Densely-overlapping regulons

Shen-Orr *et al.*[2] introduced the *Densely-overlapping regulons* (DOR) motif which can be thought of as an incomplete MIM, from which some parent–child links are absent. They used an indirect method to detect DORs in the network, using an average-linkage clustering method. It is unlikely that this method will enumerate all DORs. We note that our algorithm for enumerating MIMs can be extended to count DORs more systematically. For instance, as in the case of enumerating MIMs described in the section earlier, the sets $\mathcal{P}$ and $\mathcal{C}$ are first calculated such that there is an arc between every node in $\mathcal{P}$ and every node in $\mathcal{C}$. Subsequently, relaxing the constraint of having "complete" sets of edges between the two sets, $\mathcal{P}$ and $\mathcal{C}$ can be iteratively extended such that they now include nodes that share at least a certain user-defined threshold on the number of edges between the sets. Once extended, the edges between $\mathcal{P}$ and $\mathcal{C}$ can be assumed to be complete, and DORs (free of parent–parent and child–child edges) extracted using the procedure discussed earlier.

On counting FFLs

Compared with counting SIMs and MIMs, the problem of counting FFLs is much simpler. For every pair of (child) nodes that have an incoming edge from a common (source) node in the network, we count a FFL if there is an edge between the child nodes. (Note, if the edge between the child nodes is bidirectional, we then count 2 FFLs instead of 1.) This process is iterated for every (source) node of the network to arrive at the total count of FFLs in a directed network.

RESULTS

We reanalyze, using the definitions and algorithms presented here, the data on the yeast regulatory network

Table I
Corrected Statistics for Relative Occurrences of Network Motifs Across Static, Endogenous and Exogenous Processes in Saccharomyces Cerevisiae

		Endogenous		Exogenous		
Motif	Static	Cell cycle	Sporulation	Diauxic shift	DNA Damage	Stress response
Single input module (SIM)	107 (4.0%)	27 (14.5%)	27 (20.0%)	48 (19.3%)	45 (19.4%)	32 (26.7%)
Multiple input module (MIM)	**1551 (58.4%)**	56 (30.1%)	41 (30.4%)	**137 (55.0%)**	**117 (50.4%)**	**46 (38.3%)**
Feed forward loop (FFL)	997 (37.6%)	**103 (55.4%)**	**67 (49.6%)**	64 (25.7%)	70 (30.2%)	42 (35.0%)
Total	2655	186	135	249	232	120

The highest motif counts among the groups are highlighted in bold.

treated by Luscombe *et al.*[1] We find major differences from the numbers of motifs reported by Luscombe *et al.*[1] Table I contains corrected motif counts.

Changed proportions of motifs: SIMs, FFLs, and MIMs revealed major rewirings of the network under different physiological conditions.[1] Consistent applications of our definitions confirm the broad outlines of their conclusions—especially, their insight that changes in physiological state can reprogram the network—but reveal substantial differences in both absolute and relative numbers of different motifs in different states. The discrepancies arise in part from the use by Luscombe *et al.* of different definitions of the motifs, but those authors do not appear to have consistently used a single alternative definition in all phases of their calculations. The major sources of the discrepancies between our results and those they supersede are programming errors in the work of Luscombe *et al.* Use of inconsistent definitions of motifs in different phases of their calculations aggravates the problem but is not its primary cause. Furthermore, a more relaxed definition of MIMs, for example, eliminating the maximality criterion, would be expected to increase the number of MIMs counted, as discussed in the "Definitions" section, but this is not what we observe in comparing our results with those of Luscombe *et al.* Some papers count *bifan* occurrences [see Fig. 1(c)] as MIMs. However, relaxation of the maximality criterion will result in very high numbers of *bifans*. Using our definition, under static conditions of the yeast regulatory network, only ~12% of MIMs were found to be bifans. Attempts to reproduce their results using software from their web were unsuccessful leading to the conclusion that there are errors in their calculation.

In the static case, we find a total of 1551 MIMs, together containing 123 different parent nodes and 1462 different child nodes. Only 46 nodes appear as both parent and child, in different MIMs. The maximum number of parents in any MIM is 12, and the maximum number of children is 119. The "fan-out"—the ratio of number of children to number of parents in any MIM—ranges from 59.5 to 0.16. That is, there is a MIM with two parents and 119 children, and two MIMs with 12 parents and two children. We observe that the range of fan-out values is substantially reduced in nonstatic data sets. (See http://hollywood.bx.psu.edu/networks/analysis/mim-stats.html.)

Our results require a qualitative revision of the picture of the network dynamics proposed by Luscombe *et al.*[1] Using identical data sets as those of the authors, we find the following

1. In two exogenous conditions—diauxic shift and DNA damage, but not stress response—the absolute numbers and fractions of MIMs are larger than in endogenous subnetworks (cell cycle, sporulation).
2. The fraction of SIMs varies the least between exogenous and endogenous conditions.
3. Instead of MIMs being by far the rarest motif (325 MIMs, 7%) in the static case, as reported by Luscombe *et al.*,[1] we find 1551 MIMs, the most prevalent motif (58.4%). The increase in the number of MIMs in most exogenous states is coupled with a decrease in the number of FFLs. This increase may indicate a simultaneous activation of groups of genes by a set of transcription factors, allowing a combinatorial effect appropriate (and observed) in exogenous conditions.[2,3]
4. Our reported values of FFLs in the various networks are different from those reported by Luscombe *et al.*,[1] even though we use the same definition of FFL. We are confident that our values have been calculated correctly.

The full list of all FFLs, SIMs and MIMs for various networks used in this analysis is available from http://hollywood.bx.psu.edu/networks/analysis. Software we developed to enumerate all network motifs is available from authors on request.

CONCLUSIONS

We have given, for the first time, a precise and consistent definitions of SIM and MIM. We presented algorithms for counting SIMs, MIMs, and FFLs in directed networks. We reanalyzed the distribution of motifs in the Yeast regulatory network under different physiological conditions, to find major differences from previously published work.

ACKNOWLEDGMENTS

We thank P.J. Stuckey for helpful discussions on counting MIMs. We thank M.M. Babu, M.B. Gerstein, N.M. Luscombe, M. Snyder, S.A. Teichmann, and H. Yu for helpful discussion of regulatory network motifs.

REFERENCES

1. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 2004;431:308–312.
2. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Gen 2002;31:64–68.
3. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science 2002;298:824–827.
4. Alon U. An introduction to systems biology: design principles of biological circuits. London: Chapman & Hall/CRC; 2006.
5. 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 JB, Volkert TL, Praenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 2002;298:799–804.
6. Bron C, Kerbosch J. Finding all cliques of an undirected graph. Commun ACM 1973;16:575–577.