Molecular BioSystems

Cite this: DOI: 10.1039/c1mb05241a

www.rsc.org/molecularbiosystems

PAPER

Enrichment and aggregation of topological motifs are independent organizational principles of integrated interaction networks†‡

Tom Michoel,*[a] Anagha Joshi,[b] Bruno Nachtergaele[c] and Yves Van de Peer[d]

Received 16th June 2011, Accepted 3rd August 2011
DOI: 10.1039/c1mb05241a

Topological network motifs represent functional relationships within and between regulatory and protein–protein interaction networks. Enriched motifs often aggregate into self-contained units forming functional modules. Theoretical models for network evolution by duplication–divergence mechanisms and for network topology by hierarchical scale-free networks have suggested a one-to-one relation between network motif enrichment and aggregation, but this relation has never been tested quantitatively in real biological interaction networks. Here we introduce a novel method for assessing the statistical significance of network motif aggregation and for identifying clusters of overlapping network motifs. Using an integrated network of transcriptional, posttranslational and protein–protein interactions in yeast we show that network motif aggregation reflects a local modularity property which is independent of network motif enrichment. In particular our method identified novel functional network themes for a set of motifs which are not enriched yet aggregate significantly and challenges the conventional view that network motif enrichment is the most basic organizational principle of complex networks.

1 Introduction

Reconstructing the organizational principles that determine the structure and function of regulatory and protein–protein interaction networks is a key challenge of network biology. Network motifs, small subgraphs occurring significantly more often than expected by chance, have been proposed as the basic building blocks of complex networks,[1,2] including integrated networks composed of multiple types of interactions.[3–7] In transcriptional regulatory networks, network motifs are known to aggregate into larger, self-contained units.[2,8,9] This concept was extended to integrated networks and resulted in the definition of 'network themes', frequently recurring higher-level patterns of overlapping network motifs,[4] which characterize the structure of functional modules.[10] Several studies have further investigated the connection between network motif enrichment and aggregation, from a topological as well as from an evolutionary perspective. In hierarchical scale-free random networks the enrichment and aggregation of a certain class of subgraphs are intimately related to each other and to the global topological network parameters.[11] Furthermore these subgraphs tend to aggregate around network hubs.[11] A comparative phylogenetic analysis of genes within motifs has shown that they are not subject to any evolutionary pressure to preserve the motif pattern.[12] A likely reason is that the motifs aggregate and cannot be considered in isolation.[12,13] In a simple duplication–divergence model for network growth, modularity, accompanied by subgraph abundance, can appear for free without selection pressure.[14] On the other hand, in a model of evolving electronic circuits, modularity emerges only in an environment that changes itself in a modular manner, while network motif enrichment appears only if the modularly varying goal contains information-processing tasks.[15]

An important question that has not been addressed before is whether the aggregation of a motif is indeed surprising or significant, given the number of motif instances that have to fit on a network with a fixed degree distribution, and if so, how such aggregation relates to motif enrichment and whether any functional interpretation can be given to it. Here we address this question using random network ensembles which preserve the degree distribution as well as the total motif count, a new network motif aggregation statistic, and a novel algorithm for identifying clusters of overlapping motifs to assess in a quantitative way the enrichment and aggregation

[a] *Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Albertstrasse 19, 79104 Freiburg, Germany. E-mail: tom.michoel@frias.uni-freiburg.de; Fax: +49 761 203 97323; Tel: +49 761 203 97346*
[b] *Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building Hills Rd, Cambridge CB2 0XY, UK*
[c] *Department of Mathematics, University of California, Davis, One Shields Avenue, Davis, CA 95616-8366, USA*
[d] *Department of Plant Systems Biology, VIB and Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium*

† Published as part of a Molecular BioSystems themed issue on Computational Biology: Guest Editor Michael Blinov.
‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05241a

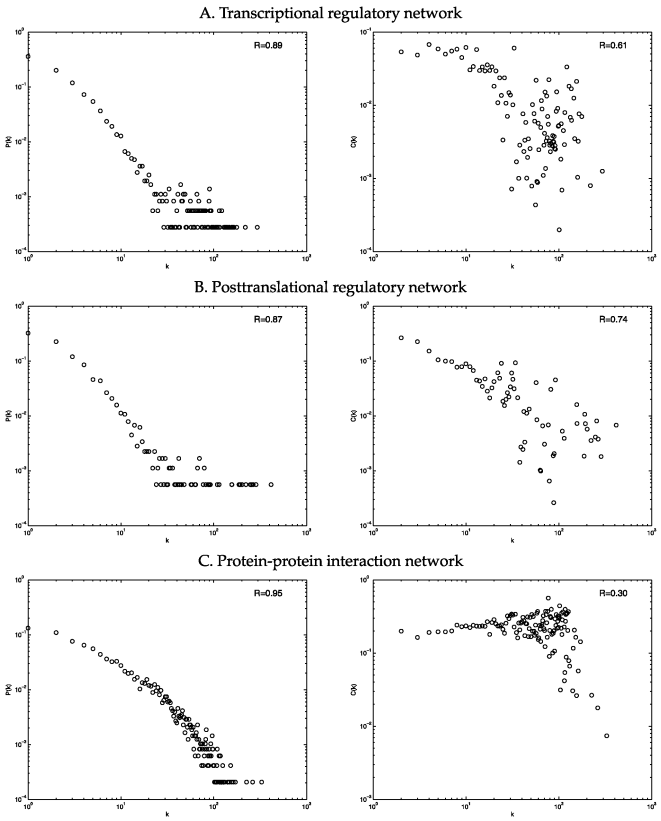

Fig. 1 Degree distribution $P(k)$ (left) and clustering coefficient $C(k)$ (right) as a function of the degree k for three molecular interaction networks in yeast. In the upper right corner of each figure, the correlation coefficient R of the data to the best power-law fit is given.

significance of all composite motifs in a network which integrates transcriptional[16] and posttranslational[5] regulatory interactions as well as physical protein–protein interactions[17] in yeast.

2 Results

2.1 Molecular interaction networks in yeast deviate from the hierarchical scale-free model

The most detailed study of the relation between network motif enrichment and aggregation to date has been done for so-called hierarchical scale-free random networks, which are characterized by a power-law degree distribution $P(k) \approx k^{-\gamma}$, where $P(k)$ is the probability for a node to have k neighbors (irrespective of edge direction), and a power-law scaling for the clustering coefficient $C(k) \approx k^{-\alpha}$ where $C(k)$ is the average clustering coefficient for a node with k neighbors.[11] Hierarchical scale-free random networks share with biological networks the property that they are organized into many small, highly connected modules that combine hierarchically into larger, less cohesive units.[11,18] The hierarchical scale-free model predicts that highly abundant network motifs always aggregate into larger motif clusters centred around network hubs in order to distribute a large number of motifs over a comparatively small number of nodes.[11]

Although Vázquez *et al.*[11] showed that several biological networks could be approximated by the hierarchical scale-free model, new data on these networks have accumulated since then. We calculated the degree distribution $P(k)$ and clustering coefficient distribution $C(k)$ for the transcriptional[16] and posttranslational[5] regulatory networks as well as for the physical protein–protein interaction network[17] in yeast (Fig. 1). For all three networks the degree distribution fits well to a power-law (correlation coefficient $R = 0.89,0.87,0.95$, respectively), with deviations mainly at the (relatively few) high-degree nodes or hubs. The clustering coefficient distribution however shows significant deviation from power-law scaling already at medium-degree nodes ($R = 0.61, 0.74, 0.30$, respectively). In other words, many medium-degree nodes in these networks have a significantly higher clustering coefficient than expected from the hierarchical scale-free model, suggesting the presence of an additional organizational level.

2.2 A network motif aggregation statistic to quantify local modularity

We hypothesized that deviations of the clustering coefficient distributions from power-law behavior are due to a local aggregation of network motifs around specific nodes which are not necessarily hubs. To quantify this aggregation of network motifs, we made the following considerations. If two networks share the same number of instances of a given motif, then the motif is more aggregating in the network where fewer nodes participate in one of the motif instances. Conversely, if two networks have the same number of nodes participating in motif instances, the motif is more aggregating in the network which has the highest number of motif instances among these nodes. We defined a network motif aggregation

statistic $\mathscr{S}$ for any three-node motif, having exactly these properties, as the ratio

$$\mathscr{S} = \frac{N}{\sqrt{n_1 n_2 n_3}} \quad (1)$$

where N is the total number of motif instances and n_1, n_2, and n_3 are the number of network nodes which participate at least once in a motif at each of the three possible motif nodes (see Methods for details).

We used this statistic to compare the real networks to randomized networks which preserve the in- and out-degree distributions as well as the total number of instances of the input motif of the real network. This random network ensemble is different from the usual one which only preserves the degree distributions. By adding the constraint to also preserve the total motif count, we ensure to assess aggregation of network motifs independent of their abundance or enrichment. We say a network exhibits *local modularity* (as opposed to hierarchical modularity) with respect to a certain motif if its aggregation statistic is significantly higher in the real network than in the randomized networks.

2.3 Feedforward loop aggregation in yeast regulatory networks is independent of enrichment

We first considered the feedforward loop (FFL) in the transcriptional and posttranslational regulatory network. The transcriptional FFL is undoubtedly the best studied network motif and its functional role and aggregation have been described in several studies.[2,4,8,9,19,20] It is strongly enriched ($P < 0.001$) in our network and also significantly aggregating ($P < 0.001$). Interestingly, the FFL is not at all enriched in the posttranslational network ($P > 0.999$), where it in fact occurs significantly less often than expected by chance. However the posttranslational FFL is strongly aggregating ($P < 0.001$). This result already indicates that network motif enrichment and aggregation are not in one-to-one relation like in the hierarchical scale free model and that the deviations of the clustering coefficient distributions from the hierarchical scale-free model in both networks (Fig. 1) are well represented by the network motif aggregation statistic.

2.4 Composite network motifs also exhibit local modularity independent of their enrichment

An additional reason why the clustering coefficient distributions may deviate from the hierarchical scale-free model is the fact that the transcriptional, posttranslational and protein–protein interaction network do not exist in isolation but are intertwined with each other. Network motifs composed of multiple interaction types represent the functional relationships between different levels of regulation in a cell.[3,4,6] Hence we examined the enrichment and aggregation of all three-node composite motifs which occur at least 100 times in the integrated network of transcriptional, posttranslational and protein–protein interactions (Fig. 2). There appears to be no strong relation between network motif enrichment and aggregation and in particular, there are several examples of motifs which are not enriched yet display significant aggregation (Fig. 2A). The Spearman rank correlation between the enrichment and aggregation Z-scores is 0.55, indicating that both properties are at best weakly correlated to each other (Fig. 2C).

2.5 An algorithm to identify local clusters of overlapping network motifs

To analyze in more detail the functional role of network motif aggregation, we developed an algorithm to identify network motif clusters. Kashtan *et al.*[8] introduced the concept of topological motif generalizations which consist of perfect motif replications along one of the motif nodes (Fig. 3A). To allow for the possibility of imperfect networks with missing interactions, we further generalized this concept and defined motif clusters as subnetworks which locally maximize the aggregation statistic $\mathscr{S}$ (*cf.* eqn (1)). In a motif cluster, each motif node i corresponds to a 'node role'[8] and is replicated into a set of cluster nodes X_i (Fig. 3B). The aggregation score of a cluster is defined as the aggregation statistic restricted to the subnetwork formed by X_1, X_2 and X_3. To find high-scoring clusters, we defined cluster membership weights for each node role, similar to spectral weights for matrices,[21] such as the PageRank[22] or hub- and authority[23] weights: a node gets a high weight in role 1, if it belongs to many motif instances together with nodes which have high weights in roles 2 and 3, and similarly for the other roles. This yields a set of multilinear equations in the membership weights for each role which are easily solved numerically. After taking a suitable threshold on the weight vectors a high-scoring cluster is obtained. The algorithm continues in an iterative fashion by removing from the network all motif instances assigned to the previous cluster and repeating the procedure until no more instances remain. Since motif instances are partitioned, nodes and edges can belong to multiple clusters. We refer to the Methods section for more details.

2.6 Comparison with Zhang *et al.*

Zhang *et al.*[4] defined network themes as classes of motif clusters based on visual inspection of composite network motifs. Our definition of local modularity on the other hand is based on the significance of the network motif aggregation statistic and provides an unbiased and rigorous method for identifying network themes. Three of the network themes discovered by Zhang *et al.*[4] pertain to the networks studied here, namely the transcriptional feedforward loop (Fig. 2A(1)), the transcriptionally coregulated interacting proteins motif (Fig. 2A(2)) and the copointing interacting transcription factors motif (Fig. 2A(3)). In all three cases our method found a highly significant aggregation Z-score, confirming the validity of our approach.

Having furthermore an automated clustering algorithm allows us to identify functional themes associated to each locally modular network motif. For instance, among the functional categories enriched in transcriptional FFL clusters, we find mainly the core processes associated with transcription such as transcriptional control, DNA binding and regulation of metabolic processes (Table S1, ESI‡), supporting the hypothesis that transcriptional FFLs play a universal information-processing role.[24] For the transcriptionally coregulated interacting proteins motif, it is usually assumed that enrichment and

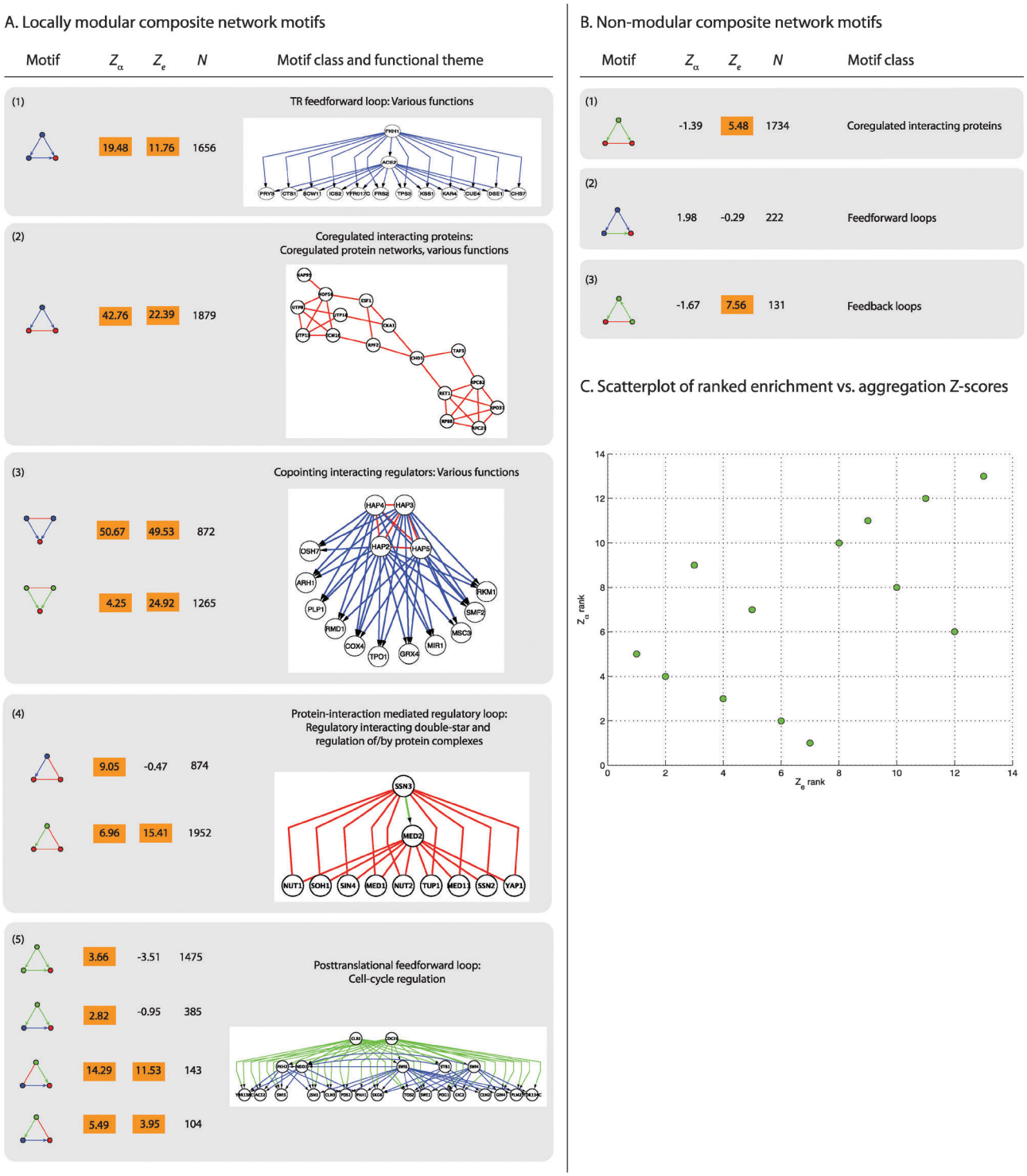

Fig. 2 Significant locally modular (A) and non-modular (B) composite network motifs in the integrated yeast network, organized by common motif classes and functional themes. Shown is for each motif the aggregation Z-score (Z_α), the enrichment Z-score (Z_e) and the number of instances (N), and for each functional theme an example of a motif cluster. Significant Z-scores ($P < 0.005$) are highlighted in orange. Interaction color legend: TR, PTL and PPI interactions, respectively, in blue, green and red. (C) Scatterplot of Z_e *vs.* Z_α ranks (in ascending order) for all motifs in panels A and B.

clustering reflects a 'regulonic complex' theme in which transcriptionally coregulated interacting proteins are often members of a protein complex.[3,4,25] We found that high-scoring coregulated protein clusters sometimes overlap with known protein complexes (Table S2, ESI‡), but more often form 'functional protein networks'[26] (Fig. 2A(2)): subnetworks of the PPI network enriched for a particular function and identified by overlaying the protein interaction network with an additional layer of information, in this case regulator–target data. Functional coregulated protein networks can be of practical interest to

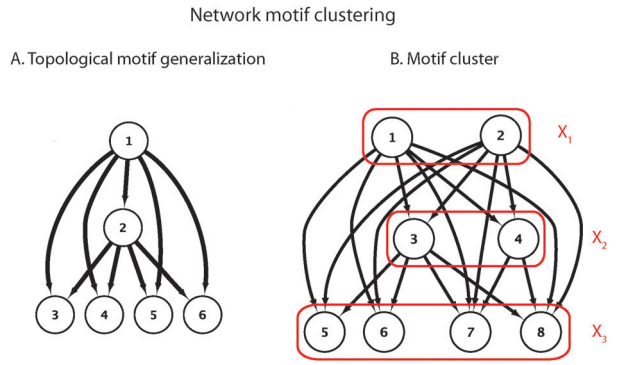

Fig. 3 (A) Example of a topological motif generalization where all possible motif instances (in this case FFL) between the nodes are present. (B) Example of a motif cluster (in this case FFL) with a high aggregation score (high number of motif instances relative to the number of nodes in X_1, X_2, and X_3).

generate detailed hypotheses for the different functions in which a particular regulator is involved. For instance ABF1 is a multifunctional global regulator, but its set of targets in the transcriptional network is only enriched for tRNA synthesis. Network motif clustering on the other hand identifies protein networks regulated by ABF1 enriched for several more categories, many of which are consistent with current knowledge, such as general DNA binding function, regulation of ribosome biosynthesis,[27] nuclear transport,[28] *etc.* (Table S3, ESI‡). Interestingly, the posttranslationally coregulated interacting proteins motif is not significantly aggregating, although it is significantly enriched (Fig. 2B(1)). This means that targets regulated by the same kinase physically interact more often than randomly selected proteins, but the resulting coregulated protein networks are not more dense than expected by chance. This result is consistent with the fact that protein complexes can be posttranslationally regulated by regulating just one instead of all of its components (see also Section 2.8).

Transcription factors often function as a complex and the binding sites for these transcription factors occur more frequently within the same promoter regions[16] leading to a 'copointing' theme.[4] The copointing interacting regulators motif is significantly aggregating at transcriptional and posttranslational levels (Fig. 2A(3)). At the transcriptional level, copointing interacting regulator pairs include well-known co-operating transcription factors like the cell cycle transcription factors SWI4–SWI6–MBP1, the ribosomal protein regulators RAP1–FHL1, the galactose response regulatory complex GAL3–GAL80, and others. At the posttranslational level, many clusters come from the multi-functional cyclin-dependent kinases (CDK) CDC28 or PHO85 complexed with one of their many cyclin activators.[29,30]

We conclude that highly abundant network motifs which are strongly enriched as well as aggregating likely have a universal information-processing role which extends across various functional categories.

2.7 Posttranslational feedforward loop aggregation reflects a cell-cycle regulation theme

Perhaps the most surprising finding in our analysis is the existence of network motifs which aggregate significantly but are not enriched (Fig. 2). We hypothesize that local modularity of a network motif without significant enrichment indicates that this motif is important for specific biological functions but does not play a universal role like the strongly enriched motifs in the previous section. Two examples of such motifs are the posttranslationally controlled feedforward loops (Fig. 2A(5)).

More than half of the posttranslational FFLs and mixed posttranslational–transcriptional FFLs belong to clusters regulated by the multi-functional CDKs CDC28 or PHO85 and these clusters are indeed often enriched for cell-cycle related functions (Tables S4 and S5, ESI‡). CDC28 is the central coordinator of the yeast cell cycle[29] and it has been shown that the mixed posttranslational–transcriptional FFLs regulated by it are important transducers between cell cycle regulatory signals and responses, using dynamical models for individual motif instances.[31] Our approach on the other hand reveals the overlapping nature of these motifs. For instance, cluster 1 (depicted in Fig. 2A(5)) contains three transcription factors functioning in G1/S transition (SWI4, SWI6, STB1), one in G2 phase (FKH2) and one in G2/M transition (NDD1). The complexity of the overlapping motif structure is further emphasized by the fact that NDD1 not only functions as a transcriptional transducer of the cell cycle signal, but also as a response target of the four other transcription factors. The target proteins in this cluster are enriched for several cell-cycle related functions (Table S5, ESI‡) and eleven of the sixteen targets are periodically expressed.[32,33] PHO85 is another CDK with a multifunctional role in cell cycle control and other processes.[30] Like CDC28 it is activated by a large family of cyclins. The transcription factors associated to clusters regulated by PHO85 contain the PHO85 substrates PHO4, GCN4 and SWI5 whose phosphorylation is important for the role of PHO85 in regulating environmental signalling response and the cell cycle.[30] This suggests that the posttranslational–transcriptional FFL plays a similar dynamical role in transducing PHO85 regulatory signals as for CDC28.[31]

The cell-cycle is a complex process and cell-cycle kinases often also interact physically with their target substrates. As a result there are two motifs involving all three interaction types (Fig. 2A(5)) which almost all overlap with mixed posttranslational–transcriptional FFLs. The aggregation significance is consistent across all four posttranslational FFLs but the enrichment is not (Fig. 2A(5)). This is not in contradiction with the previous result that enriched locally modular motifs play a universal role, since the vast majority of triple-interaction motifs involve cell-cycle regulators, *i.e.* 'universal' always refers to the network at hand.

In summary, we can say that the posttranslational regulatory network exhibits an overall lack of feedforward loops, presumably because it operates on a much shorter timescale than the transcriptional regulatory network to elicit fast information-processing responses, typically in the form of signaling cascades.[24] Posttranslational feedforward loops (pure as well as composite) do seem to play an important role however in regulation of the cell-cycle, and this 'local' role is reflected in a significant aggregation of these motifs around the core CDKs CDC28 and PHO85.

2.8 Transcriptional and posttranslational protein interaction-mediated regulatory loop aggregation reflects a regulatory protein complex theme

Another motif which is not enriched yet displays significant aggregation is the protein-interaction mediated transcriptional regulatory loop (Fig. 2A(4)), a circuit that is thought to serve for feedback mechanisms between a regulator–target pair *via* a common partner in the protein interaction network.[34] In the equivalent posttranslational motif all interactions can occur simultaneously and its proposed function is that of a 'scaffold motif' where the biochemical interaction between the regulator and its target substrate is enabled by the common interactor.[5] A natural cluster generalization of such feedback or scaffold circuits is a 'regulonic star', where multiple targets of a regulator ('spokes') interact with the same feedback or scaffold mediator ('hub'). Our algorithm identified several such modules as high-scoring motif clusters (Table S6 and S7, ESI‡). For instance transcriptional cluster 11 consists of ABF1, a DNA binding protein that regulates multiple nuclear events, regulating a set of ten nuclear transport genes which all interact with PSE1, a nuclear transport receptor which also interacts with ABF1. A link between ABF1 and the nuclear transport machinery *via* PSE1 is known.[28] Transcriptional cluster 14 has HSP82 as the hub protein. HSP82 is one of two yeast genes encoding for HSP90, a protein folding chaperone which plays a central role in various aspects of cellular signaling.[35,36] Binding of HSP90 to HAP1, the regulator of cluster 14, is necessary for heme activation of HAP1.[37] This cluster may represent a feedback mechanism since TAH1, a cofactor of HSP90,[36] is one of its spoke proteins. Posttranslational cluster 43 is an example of a scaffolding regulonic star. It consists of the mitotic B-type cyclin CLB2 which phosphorylates nine proteins involved in budding, cell polarity and filament formation, which all interact with NAP1, a protein which is known to interact with and facilitate the function of CLB2.[38]

We also found a relation between the protein-interaction mediated regulatory loops and protein complexes in the form of a 'regulatory interacting (RI) double-star' cluster type, consisting of one or a few regulator–target pairs which share a common set of partners in the protein interaction network. Usually the spoke proteins in such a RI double-star mutually interact and form the components of a protein complex, often together with the hub protein (Tables S8 and S9, ESI‡). For instance, in posttranslational cluster 32, SSN3 (also called SRB10) phosphorylates MED2, a component of the RNA polymerase II Mediator complex, and both interact with eight other Mediator components and the transcription factor YAP1 (depicted in Fig. 2A(4)). It is known that posttranslational modification of Mediator components affects its function, and that SRB10 is part of a module whose binding to the Mediator complex determines if Mediator can associate with pol II or not.[39] YAP1, a bZIP transcription factor required for oxidative stress tolerance, is related to the Mediator complex *via* a transcriptional RI double-star in which 25 components of the Mediator complex interact with YAP1 and SRB6 and SRB7, two other components of the Mediator complex that are transcriptionally regulated by YAP1. The Mediator complex acts as a bridge between gene-specific transcription factors and the basal pol II transcription machinery[39] and mutants of the general transcription factor TFIIA are unable to grow in conditions that require the oxidative stress response.[40] Another example of a transcriptional RI double-star is transcriptional cluster 20 where HAP4 regulates GCN4 and both interact with five components of the SWI/SNF complex, which regulates transcription by nucleosome remodeling. HAP4 and GCN4 are two transcriptional activators which target SWI/SNF to appropriate promoters.[41]

The protein complexes which appear in transcriptional regulatory interacting double-stars are all regulatory complexes involved in the different steps from chromatin remodelling to transcription, translation and posttranslational control (Table S8, ESI‡). In cases where the hub protein also belongs to the complex, this suggests that the RI double-star acts like a two-node feedback loop in which the transcription factor regulates the complex by regulating one or a few of its components, and in turn the complex regulates the transcription factor by interacting with it at the protein level. A composite feedback mechanism using a slow (transcriptional) and fast (protein–protein interaction) timescale is known to enhance stability around a steady state.[24] For the posttranslational motif on the other hand, RI double-star clusters reflect regulation of the protein complex, and hence we find a much broader range of protein complexes besides regulatory complexes (Table S9, ESI‡). This more universal role is again consistent with the fact that the posttranslational motif is strongly enriched but the transcriptional is not. The interpretation of a posttranslational RI double-star cluster is that the kinase and protein complex form a two-component loop where the complex acts as a scaffold protein for its own regulation.

Regulatory interacting double-star clustering of protein-interaction mediated regulatory loops induces a higher-level, global map of protein complex regulation. In Fig. 4 we considered all protein complexes which overlap significantly with RI double-star clusters with at least three (transcriptional) or four (posttranslational) spoke proteins. This map shows a high amount of two-component regulator–complex feedback loops, with a central role for the Mediator complex and the nucleosomal proteins. Interestingly, transcriptional and posttranslational regulations are heavily intertwined in this map. Some complexes (Mediator, small ribosomal subunit, nucleosomal proteins) are regulated by transcriptional as well as posttranslational RI double stars, while others (Srb10p, SLIK) play a feedback or scaffolding role for transcriptional as well as posttranslational regulatory interactions. Fig. 4 provides a novel kind of coarse grained integrated network representation which complements previous thematic maps of compensatory and regulonic complexes.[4]

The protein-interaction mediated transcriptional regulatory loop has previously been found enriched[4] or not enriched[3] in different datasets. We calculated the aggregation statistic in these two datasets as well as for a network of literature-curated protein–protein interactions,[42] in which the motif is also enriched, and found in all cases a consistent statistically significant aggregation (data not shown). The protein interaction networks where the motif is enriched are targeted towards co-complex interactions, while the networks where it is not enriched also contain interactions derived from

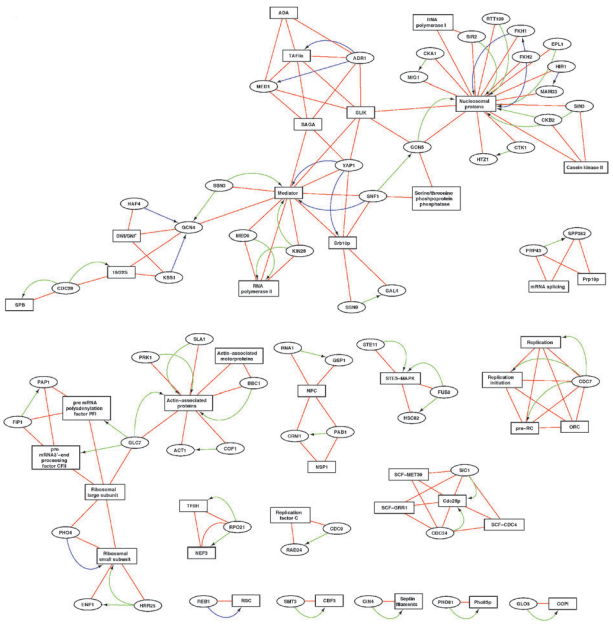

Fig. 4 Global map of protein complex regulation through transcriptional and posttranslational regulatory interacting double-star clusters. Oval nodes are proteins, rectangular nodes are protein complexes which overlap significantly ($P < 10^{-4}$) with a RI double-star cluster. Two-node mixed feedback loops indicate that the target hub protein belongs to the same complex as the spoke nodes. This figure is based on the data in ESI‡, Tables S8 and S9. The interaction colors are the same as in Fig. 2.

yeast two-hybrid studies. Thus we find that network motif aggregation is a more robust property than network motif enrichment and that the enrichment in co-complex based networks can be explained by the relation between the protein-interaction mediated transcriptional regulatory loop and protein complexes as detailed in the previous paragraphs. Our results also highlight the advantage of using an unbiased clustering approach to identify network themes, since Zhang *et al.*[4] did consider the transcriptional protein interaction-mediated loop but found no theme or cluster generalization for it.

3 Conclusions

Network motifs form the basic building blocks of complex networks and previous studies have suggested an intimate relation between the enrichment of a motif and its tendency to aggregate into functional modules. Evolutionary studies hypothesized that network motif clusters have evolved by simple duplication–divergence mechanisms, concluding that the abundance of a motif is merely a by-product of the emergence of these aggregated motif structures. On the other hand, a purely topological analysis in hierarchical scale-free random networks has shown that the only way to distribute a large number of motifs over a comparatively small number of nodes is to aggregate the motifs around the network hubs, indicating that motif aggregation follows from motif enrichment.

Here we introduced a novel method for assessing in a quantitative way the statistical significance of network motif aggregation. Using an integrated network of transcriptional, posttranslational and protein–protein interactions in yeast, we showed that our method produces results which are consistent with deviations of the clustering coefficient distribution from the hierarchical scale-free model and correctly recapitulate previous findings. We showed furthermore that significant aggregation reflects local modularity of network motifs around nodes which are not necessarily hubs in the total network but play a core role in specific biological processes. Using a novel network motif clustering algorithm we found that if in addition to aggregate significantly, a motif is also enriched, it likely plays a universal functional role across many biological processes. If the motif aggregates significantly without enrichment, it is likely specific for one or a few processes. We identified novel functional network themes for such motifs, like a cell-cycle regulatory theme for posttranslationally controlled feedforward loops and a regulation of or by protein complexes for protein-interaction mediated regulatory loops.

Our results show that network motif aggregation is an important organizational principle of molecular interaction networks which is independent of network motif enrichment and in particular is more robust against using heterogeneous experimental sources of interaction data and less sensitive to the incompleteness of all current interaction networks. We hypothesize that network motif aggregation is the more fundamental organizational principle and that network motif enrichment may follow from aggregation if the local modularity property extends across a sufficiently large fraction of biological processes represented in the network at hand. Future work to support this hypothesis should be directed towards comparative analyses to confirm these results across multiple organisms and towards improving the current duplication–divergence models for network growth and hierarchical scale-free models for network topology in order to reproduce and understand the evolutionary origin of the complex modular organization observed in the integrated yeast interaction network.

4 Methods

4.1 Network data

The protein–protein interaction network (36391 interactions between 4847 proteins) was extracted from the BioGRID database[17] (http://www.thebiogrid.org). The transcriptional regulatory network (11 373 interactions between 198 transcription factors and 3535 target genes) was obtained from Harbison *et al.*[16] (http://fraenkel.mit.edu/Harbison/), using a *P*-value cutoff of 0.005. The posttranslational regulatory network (5630 interactions between 264 regulators and 1653 target proteins) was extracted from the BioGRID database (interactions annotated as 'biochemical activity').[17] The bulk of this network (4621 interactions) comes from the phosphorylation network of Ptacek *et al.*[5] (http://networks.gersteinlab.org/phosphorylome/). Self-interactions were not kept in any of the networks.

4.2 Network motif aggregation statistic

For a given 3-node motif we choose a particular labeling of its nodes and define its aggregation statistic as

$$\mathscr{S} = \frac{N}{\sqrt{n_1 n_2 n_3}} \qquad (2)$$

where N is the total number of motif instances and n_1, n_2 and n_3 are the number of nodes which participate at least once in a motif in node roles 1, 2 and 3, respectively. $\mathscr{S}$ has the intuitive properties that it is higher (more aggregation) in a network with a fixed number of motif instances distributed over a smaller number of nodes or in a network with more motif instances distributed over a fixed number of nodes. We have $N \leq n_1 n_2 n_3$, and the maximum is attained for perfect topological motif generalizations as in Fig. 3A where all possible motif instances are indeed present. The square root in eqn (2) ensures that $\mathscr{S}$ will be higher for bigger topological motif generalizations as well as for larger sets of nodes with a significant number of motif instances between them (as in Fig. 3B), and is thus suitable to measure motif aggregation in noisy interaction data with potentially a large number of missing interactions.

4.3 Local network motif aggregation score

We define network motif clusters as subnetworks which locally maximize the network motif aggregation statistic. To make this precise, for a given 3-node input motif, we define a 3-dimensional motif array T by

$$T_{ijk} = \begin{cases} 1 & \text{if there exists a motif instance between } i, j, k \\ 0 & \text{otherwise} \end{cases}$$

where (i, j, k) is any triple of nodes and the order of the indices corresponds to a particular labeling of the nodes in the motif. T can be constructed from the adjacency matrices of the networks defining the motif. For instance, the motif array for the feedforward loop in a directed network with adjacency matrix A is given by

$$T_{ijk} = A_{ij}A_{jk}A_{ik}.$$

A motif cluster is now defined by three sets of nodes (X_1, X_2, X_3) (*cf.* Fig. 3) and its aggregation score can be written as

$$\mathscr{S}(X_1, X_2, X_3) = \frac{\sum_{i\in X_1, j\in X_2, k\in X_3} T_{ijk}}{\sqrt{|X_1||X_2||X_3|}}, \qquad (3)$$

where $|X|$ denotes the number of nodes in X.

4.4 Network motif clustering algorithm

We want to find (X_1, X_2, X_3) which maximize the local aggregation score. To this end, we first find the best rank-1 approximation to T,[43] *i.e.* find real-valued vectors (u, v, w) maximizing

$$\mathscr{R}(u, v, w) = \frac{\sum_{ijk} T_{ijk}u_i v_j w_k}{\| u \| \| v \| \| w \|}, \qquad (4)$$

where $\| u \| = \sqrt{\sum_i u_i^2}$ is the length of u. These maximizing vectors can be found efficiently by a multilinear power method.[43] For a set of nodes X we define an index vector u_X by

$$u_{X,i} = \begin{cases} 1 & \text{if } i \in X \\ 0 & \text{otherwise} \end{cases}$$

such that $\mathscr{S}(X_1, X_2, X_3) = \mathscr{R}(u_{X_1}, u_{X_2}, u_{X_3})$. This property is used to prove that for any X_1, X_2, X_3

$$|\mathscr{S}_{\max} - \mathscr{S}(X_1, X_2, X_3)| \leq \sqrt{2}\mathscr{R}_{\max}(\| u - u_{X_1} \| + \| v - u_{X_2} \| + \| w - u_{X_3} \|) \qquad (5)$$

where $\mathscr{S}_{\max}$ is the (unknown) maximal value of $\mathscr{S}$ over all possible node sets, $\mathscr{R}_{\max}$ is the (known) maximal value of $\mathscr{R}$ over all real-valued vectors, (u, v, w) is the (known) best rank-1 approximation to T, and all vectors on the r.h.s. are normalized to length 1. Using the fact that (u, v, w) have nonnegative entries, it is trivial to find (X_1, X_2, X_3) which minimize, respectively, $\|u - u_{X_1}\|$, $\|v - u_{X_2}\|$ and $\|w - u_{X_3}\|$, or equivalently, maximize $\langle u, u_{X_1}\rangle$, $\langle v, u_{X_2}\rangle$ and $\langle w, u_{X_3}\rangle$, where

$$\langle u, u_X\rangle = \frac{1}{\sqrt{|X|}}\sum_{i\in X} u_i \qquad (6)$$

is the overlap between u and u_X. By eqn (5), these (X_1, X_2, X_3) are the best possible approximation to the highest scoring motif cluster, given our knowledge of the best rank-1 approximation to T. The r.h.s of eqn (5) gives a precise estimate on the quality of this approximation. Next we remove from the motif array T all entries corresponding to the motif instances in the highest scoring motif cluster. The procedure is repeated for this truncated motif array and iterated until no more non-zero entries remain, thus obtaining a partition of all motif instances into high-scoring motif clusters.

For symmetric motifs (such as *e.g.* the coregulated interacting proteins motif, Fig. 2A(2)), the motif array is symmetric for interchanging two indices, $T_{ijk} = T_{ikj}$ for all triples (i, j, k). The algorithm proceeds in the same way as before but ensures that a symmetric maximizer of eqn (4) is found with $v = w$, resulting in symmetric clusters with $X_2 = X_3$.

4.5 Network randomization algorithms

To assess network motif enrichment we generated random networks with the same incoming and outgoing degree distributions as the real networks as follows. A directed network with N_e edges can be represented by two index vectors $\{I, J\}$ of length N_e such that the kth edge points from node I_k to node J_k. We first generated a random permutation $\pi(J)$ of the indices in J. The network represented by $\{I, \pi(J)\}$ automatically has the same in- and out-degrees as the original network, except there may be some unwanted self-interactions, *i.e.* indices k where $I_k = \pi(J)_k$. We swap these $\pi(J)_k$ with a randomly chosen entry $\pi(J)_l$ until we obtain a corrected permutation $\pi_c(J)$ without any self-interactions and a corresponding randomized network $\{I, \pi_c(J)\}$. Undirected networks are treated in the same way except we choose $I_k < J_k$ for every k. After randomly permuting J and correcting for self-interactions as before, we swap columns for every $I_k > \pi(J)_k$ and correct for any duplicate edges by again randomly swapping entries in $\pi(J)$. This randomization strategy is easier to implement and runs faster than the conventional single edge swapping strategies.

To assess network motif aggregation we generated random networks with the same incoming and outgoing degree distributions as well as the same total motif count as the real networks as follows. First we generate random networks with the same in- and out-degrees as described in the previous paragraph. If the total number of motif instances is smaller in the random network than the real one, we generate the list of all 'incomplete' motif instances, triplets of nodes with two motif interactions present and one absent (Fig. 5A). We randomly select one of these incomplete motifs. For the two

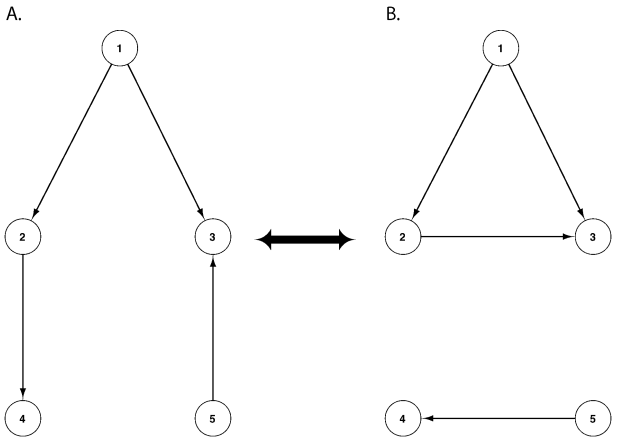

Fig. 5 Example of edge swapping operations to increase (A → B) or decrease (B → A) the number of FFLs in a random network while keeping the in- and out-degree distributions constant.

nodes with a missing edge between them, we randomly select an incoming, resp. outgoing edge. We then swap the endpoints of these edges to 'close' the incomplete motif while preserving the degree distributions (Fig. 5, A → B). If the total number of motif instances is larger in the random network than the real one, we randomly select a motif instance and edge in that motif (Fig. 5B). We randomly select another edge not belonging to any motif instance and swap the endpoints of the motif edge with this edge to 'open' a motif instance (Fig. 5, B → A). We iterate between closing or opening motif instances until the total number of motif instances is equal between the random and real network.

4.6 Network motif enrichment and aggregation significance

To compute network motif enrichment significance, we generated 1000 random networks with the same in- and out-degree distributions as the real transcriptional, posttranslational and protein–protein interaction networks. The enrichment *P*-value is defined as the fraction of random networks having at least the same number of motif instances as the real networks, and the *Z*-score is defined as

$$Z = \frac{N - \mu}{\sigma}$$

where N is the number of motif instances in the real network and μ, resp. σ, is the mean, resp. standard deviation, of the number of motif instances in the random network ensemble. To compute network motif aggregation significance, we generated for each input motif 1000 random networks with the same in- and out-degree distributions as well as the same total motif count as the transcriptional, posttranslational and protein–protein interaction networks. The aggregation *P*-value is defined as the fraction of random networks having at least the same aggregation statistic as the real network, and the *Z*-score is defined as in the previous paragraph. Notice that enrichment can be computed for all composite motifs using a single ensemble of integrated random networks. On the other hand, to compute aggregation, a separate random network ensemble has to be generated for each input motif.

4.7 Software

A Network Motif Clustering Toolbox containing an implementation of the network motif clustering algorithm as well as functions to generate random networks and compute network motif enrichment and aggregation significance is freely available for academic purposes, including source code, from http://omics.frias.uni-freiburg.de/software/. The toolbox operates under both Matlab (http://www.mathworks.com) and Octave (http://www.gnu.org/software/octave).

Acknowledgements

We thank Eric Bonnet and Vanessa Vermeirssen for discussions and testing the software. TM wishes to thank the Department of Mathematics of the University of California, Davis, for warm hospitality during visits when part of this work was performed. The work of TM, AJ and YVdP was supported in part by IWT (SBO-BioFrame) and IUAP P6/25 (BioMaGNet). The work of BN was supported in part by the National Science Foundation under grant DMS-1009502.

References

1 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**, 824–827.
2 S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.*, 2002, **31**, 64–68.
3 E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon and H. Margalit, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 5934–5939.
4 L. V. Zhang, O. D. King, S. L. Wong, D. S. Goldberg, A. H. Y. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone and F. P. Roth, *J. Biol.*, 2005, **4**, 6.
5 J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. McCartney, M. Schmidt, N. Rachidi, S. Lee, A. Mah, L. Meng, M. Stark, D. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. Predki and M. Snyder, *Nature*, 2005, **438**, 679–684.
6 H. Yu, Y. Xia, V. Trifonov and M. Gerstein, *GenomeBiology*, 2006, **7**, R55.
7 D. Fiedler, H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, A. C. Silva, M. Shales, S. R. Collins, S. van Wageningen, P. Kemmeren, F. C. P. Holstege, J. S. Weissman, M.-C. Keogh, D. Koller, K. M. Shokat and N. J. Krogan, *Cell*, 2009, **136**, 952–963.
8 N. Kashtan, S. Itzkovitz, R. Milo and U. Alon, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2004, **70**, 031909.
9 R. Dobrin, Q. K. Beg, A.-L. Barabási and Z. N. Oltvai, *BMC Bioinformatics*, 2004, **5**, 10.
10 L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, *Nature*, 1999, **402**, C47–C52.
11 A. Vázquez, R. Dobrin, D. Sergi, J.-P. Eckman, Z. N. Oltvai and A.-L. Barabási, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 17940–17945.
12 A. Mazurie, S. Bottani and M. Vergassola, *GenomeBiology*, 2005, **6**, R35.
13 R. V. Solé and S. Valverde, *Trends Ecol. Evol.*, 2006, **21**, 419–422.
14 R. V. Solé and S. Valverde, *J. R. Soc. Interface*, 2008, **5**, 129–133.
15 N. Kashtan and U. Alon, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 13773–13778.
16 C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel and R. A. Young, *Nature*, 2004, **431**, 99–104.

17 C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res.*, 2006, **34**, D535–D539.
18 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, *Science*, 2002, **297**, 1551–1555.
19 T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. Benjamin Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, *Science*, 2002, **298**, 799–804.
20 S. Kalir and U. Alon, *Cell*, 2004, **117**, 713–720.
21 M. E. J. Newman, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2006, **74**, 036104.
22 S. Brin and L. Page, *Comput. Networks*, 1998, **30**, 107–117.
23 J. M. Kleinberg, *J. Assoc. Comput. Mach.*, 1999, **46**, 604–632.
24 U. Alon, *An introduction to systems biology: design principles of biological circuits*, Chapman & Hall/CRC, 2007.
25 K. Tan, T. Shlomi, H. Feizi, T. Ideker and R. Sharan, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 1283–1288.
26 N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, E. Ruppin and R. Sharan, *Mol. Syst. Biol.*, 2009, **5**, 248.
27 R. J. Planta, P. M. Gonçalves and W. H. Mager, *Biochem. Cell Biol.*, 1995, **73**, 825–834.
28 C. M. Loch, N. Mosammaparast, T. Miyake, L. F. Pemberton and R. Li, *Traffic*, 2004, **5**, 925.
29 M. D. Mendenhall and A. E. Hodge, *Microbiol. Mol. Biol. Rev.*, 1998, **62**, 1191–1243.
30 D. Huang, H. Friesen and B. Andrews, *Mol. Microbiol.*, 2007, **66**, 303–314.
31 A. Csikász-Nagy, O. Kapuy, A. Tóth, C. Pál, L. J. Jensen, F. Uhlman, J. J. Tyson and B. Novák, *Mol. Syst. Biol.*, 2009, **5**, 236.
32 P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, *Mol. Biol. Cell*, 1998, **9**, 3273–3297.
33 U. de Lichtenberg, R. Wernersson, T. S. Jensen, H. B. Nielsen, A. Fausbøll, P. Schmidt, F. B. Hanse, S. Knudsen and S. Brunak, *Yeast*, 2005, **22**, 1191–1201.
34 E. Yeger-Lotem and H. Margalit, *Nucleic Acids Res.*, 2003, **31**, 6053–6061.
35 A. J. McClellan, Y. Xia, A. M. Deutschbauer, R. W. Davis, M. Gerstein and J. Frydman, *Cell*, 2007, **131**, 121–135.
36 R. Zhao, M. Davey, Y.-C. Hsu, P. Kaplanek, A. Tong, A. B. Parsons, N. Krogan, G. Cagney, D. Mai, J. Greenblatt, C. Boone, A. Emili and W. A. Houry, *Cell*, 2005, **120**, 715–727.
37 H. C. Lee, T. Hon, C. Lan and L. Zhang, *Mol. Cell. Biol.*, 2003, **23**, 5857–5866.
38 D. R. Kellog and A. W. Murray, *J. Cell Biol.*, 1995, **130**, 675–685.
39 S. Björklund and C. M. Gustafsson, *Trends Biochem. Sci.*, 2005, **30**, 240–244.
40 S. M. Kraemer, D. A. Goldstrohm, A. Berger, S. Hankey, S. A. Rovinsky, W. S. Moye-Rowley and L. A. Stargell, *Eukaryotic Cell*, 2006, **5**, 1081–1090.
41 K. E. Neely, A. H. Hassan, A. E. Wallberg, D. J. Steger, B. R. Cairns, A. P. H. Wright and J. L. Workman, *Mol. Cell*, 1999, **4**, 649–655.
42 T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada and M. Tyers, *J. Biol.*, 2006, **5**, 11.
43 L. De Lathauwer, B. De Moor and J. Vandewalle, *SIAM J. Matrix Anal. Appl.*, 2000, **21**, 1324–1342.

