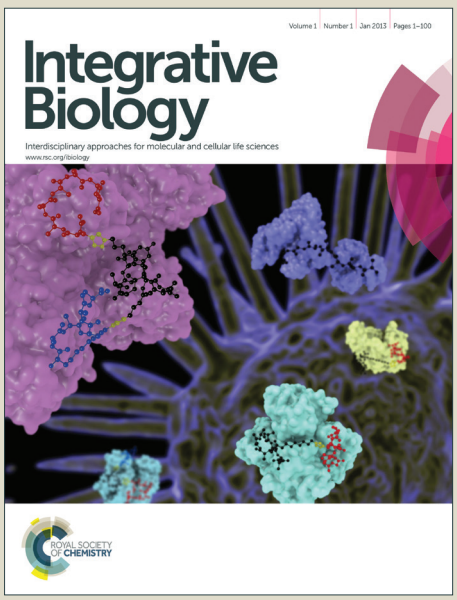
Integrative Biology

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/ibiology

Insight box statement:

The components of living cells interact with each other forming biological networks. The topology of a biological network determines to a large extent its dynamic properties and modes of operation. The controllability of a transcriptional regulatory network can be interpreted as the ability of the cell to control the expression of genes based on control by some transcription factors responding to environmental cues. Here we show that the controllability is a function of the topology and the complexity of the system. Internal loops in the network increase the controllability, but it may cause instability of the system. Thus, there seems to be a trade-off between controllability and stability of regulatory networks.

Controllability analysis of transcriptional regulatory networks reveals circular control patterns among transcription factors

Tobias Österlund[1], Sergio Bordel[1], Jens Nielsen[1,2]

[1] Novo Nordisk Foundation Center for Biosustainability, Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-41296 Göteborg, Sweden

[2] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Fremtidsvej 3, DK-2970 Hørsholm

Abstract:

Transcriptional regulation is the most committed type of regulation in living cells where transcription factors (TFs) control the expression of their target genes and TF expression is controlled by other TFs forming complex transcriptional regulatory networks that can be highly interconnected.

Here we analyze the topology and organization of nine transcriptional regulatory networks for *E.coli*, yeast, mouse and human, and we evaluate how the structure of these networks influences two of their key properties, namely controllability and stability. We calculate the controllability for each network as a measure of the organization and interconnectivity of the network. We find that the number of driver nodes n_D needed to control the whole network is 64% of the TFs in the *E.coli* transcriptional regulatory network in contrast to only 17% for the yeast network, 4% for the mouse network and 8% for the human network. The high controllability (low number of drivers needed to control the system) in yeast, mouse and human is due to the presence of internal loops in their regulatory networks where the TFs regulate each other in a circular fashion. We refer to these internal loops as circular control motifs (CCM). The *E.coli* transcriptional regulatory network, which does not have any CCMs, shows a hierarchical structure of the transcriptional regulatory network in contrast to the eukaryal networks. The presence of CCMs also has influence on the stability of these networks, as the presence of cycles can be associated with potential unstable steady-states

where even small changes in binding affinities can cause dramatic rearrangements of the state of the network.

Introduction

The number of biological network reconstructions for model organisms like *S.cerevisiae* and *E.coli* has increased dramatically during recent years and include many types of networks, e.g. signaling networks, protein interaction networks[1] and metabolic networks[2]. For the model organisms *S.cerevisiae* and *E.coli* there is a compendium of genome-scale metabolic network reconstructions (GENREs) and genome-scale metabolic models (GEMs) available, which have been used and applied for several different purposes[3, 4]. Even though these models have shown excellent capabilities in predicting different phenotypes, they do have limitations and false predictions is generally due to missing information about regulation of the metabolism [5]. Attempts to incorporate transcriptional regulation of the metabolic genes into FBA simulations have been done both for *E.coli*[5, 6] and *S.cerevisiae* [7]. Transcriptional regulation is condition dependent in the sense that most transcription factors (TFs) bind and recognize specific sequence motifs and a majority of TFs appear to regulate transcription only at specific growth conditions or under specific environmental perturbations[8, 9]. By integrating information about transcriptional regulation to the genome-scale metabolic models we would improve the ability of the model to predict a phenotype from a given genotype. In order to be able to incorporate regulatory information in the models we need to understand more about the organization and structure of transcriptional regulatory networks, as well as how the network behaves under different conditions. The organization of transcriptional regulatory networks has also an impact on the genotype to phenotype relationship in complex diseases as reviewed by Vidal et al. [10]

The first step towards understanding the regulation of biological processes on a global (i.e. genomic) scale is to reconstruct the transcriptional regulatory network (TRN). For *S.cerevisiae*, TF-DNA interactions have been characterized by ChIP-chip experiments and then been used to construct the yeast TRN[9, 11, 12]. In order to model the transcriptional regulation different approaches have been taken. The TRN can be represented as a Boolean model where the TFs are either on or off (1 or 0) based on the activity of other transcription factors and environmental factors. The Boolean modeling approach is implemented in the rFBA framework[5] where the states of the metabolic genes depend on the states of the controlling TFs. Another approach is probabilistic regulation of metabolism (PROM)[13] where

the probability of TF regulation can be estimated for each TF-target gene pair by counting if the TF and the target gene are expressed in a large number of transcriptome experiments, e.g. microarrays. The PROM method does not use Boolean logic to describe the probability of TF binding; instead the probability is continuous between 0 and 1. However it requires setting a threshold value for a gene to be expressed or not in order to estimate the probabilities of TF binding.

Human and mouse transcriptional regulation is even more complex than for *E.coli* and yeast. The human ENCODE project[14] aims to characterize and map functional elements of the human genome, including cis-regulatory elements and non-coding RNAs etc. Two databases, Cscan[15] and Chip Enrichment Analysis (ChEA)[16] have re-analyzed part of the ChIP-seq data from the ENCODE project and from other publications for human and mouse.

Here we constructed nine different transcriptional regulatory networks from different ChIP-chip and ChIP-seq experiments for *E.coli, S.cerevisiae*, human and mouse. These networks were analyzed in terms of organization, topology and network structure in order to get increased understanding about the transcriptional regulation in these organisms, and how the TRN architecture differs between different species. In order to do this we analyzed each of the networks in terms of network controllability[17] and stability, and calculated how many driver nodes are needed to control the system. For *S.cerevisiae* we also identified TFs that respond to environmental cues by analyzing microarray data from several chemostat studies where the environment was tightly controlled. By controlling the TFs that respond to the environment and calculate how many other TFs in the network that also can be controlled we obtained an understanding of the condition-specific behavior of the yeast transcriptional regulatory network.

Materials and methods

Generating TF-TF regulatory networks

The nine transcriptional regulatory networks used in this study were derived using different ChIP-chip and ChIP-seq datasets as a starting point. The three *S. cerevisiae* networks were derived from the Yeastract database 2011-10[18], the Harbison TF-DNA interaction data with binding p-value < 0.001 and with binding p-value <0.005[9]. The TRN from Yeastract includes TF-gene interactions both with direct evidence (ChIP-chip) and indirect evidence (the gene was transcriptionally changed in a TF knockout). The *E.coli* transcriptional regulatory

network was derived from the *E.coli* Regulon DB version 8.2[19]. For mouse and human TF-gene interactions were collected from the Chip enrichment analysis database (ChEA)[16]. Version 1 of the ChEA database was downloaded from http://amp.pharm.mssm.edu/chea in October 2013. Version 2 of the database was downloaded from http://amp.pharm.mssm.edu/ChEA2/ in November 2013. For human we also downloaded TF-gene interactions from the Cscan database[15]. For the controllability analysis we considered only TF-TF interactions, so all non-TF genes were filtered out. For visualization of the networks we used Cytoscape version 3.0.1. The hierarchical structures of the networks were obtained by choosing the Hierarchical layout view in Cytoscape. The simulated scale-free and random networks were constructed using the igraph R-package[20] using the Barabasi-Albert[21] and Erdos-Renyi[22] models for network growth. The simulated networks all had 100 nodes and a varying number of edges to simulate networks with different average degree.

Controllability analysis

The concept of controllability of complex networks was introduced by Liu et al. 2011[17]. The number of driver nodes n_D is defined as the minimum number of nodes that need to be controlled as input to the system to control 100 % of the network. This number is obtained from the maximum number of matched nodes in the network when solving the controllability equation. The maximum matching path is defined in Liu et al. 2011[17] as the maximum set of links in the networks that do not share start or end nodes, i.e. one path that can control all output nodes from all input nodes. Here we used linear programming to retrieve the number of maximum matching nodes. First, the TF-TF interaction network was converted to an n x m matrix, A, where the n rows represents TF nodes and the m columns represent TF-TF interactions (connections). A connection has the value 1 for the TFs that are connected and 0 for the TFs that are not involved in this connection. The following linear program was applied to obtain the maximum matching path, i.e. the longest non-overlapping path in the network, connecting the input nodes with the output nodes:

Maximize: $\sum_{i=1}^{m} x_i$,

Subject to: $Ax = b, x_i \in \{0,1\}$ for i=1,2…,m (1)

Where $x_1, x_2, \ldots x_m$ are the TF-TF interactions (connections) in the network and b is the vector of input signals which is set to 1 for one of the nodes and 0 for all other nodes in the network. The maximal number of controlled TFs, n_c is calculated as the number of nodes (TFs) that are included in the maximum matching path. The network was constructed and analyzed using

the Raven toolbox for Matlab[23]. We used the Raven function *getAllSubGraphs*() to calculate the number of internal loops (circular control motifs) and remove the loops that was not covered in the maximum matching paths, i.e. loops without any input or output nodes.

Integrated analysis of transcriptome data

The raw data files (CEL-files) for the selected microarray studies[24-35] were downloaded from Gene expression omnibus (GEO) and ArrayExpress using the accession numbers given in the papers. All 233 microarrays used the Affymetrix yeast 98 platform which made it possible to normalize all data together. The data were normalized in R using Plier normalization with only perfect match probes. Metadata for each experiment were collected and used to construct the regression model described in Equation 1. The regression model was implemented in R and ANOVA p-values were calculated for each gene and for each of the coefficients $\beta_1,\ldots,\beta_5$ where the null hypothesis for each gene is that $\beta_i=0$, and the alternative hypothesis is that $\beta_i\neq 0$ for i=1,…,5. The p-values were corrected for multiple testing using Benjamini and Hochbergs method (FDR).

A hypergeometric enrichment test was applied to identify TFs with over-represented significantly changed target genes. The TF-gene interactions were taken from the Yeastract TRN [18]. For the factors oxygen availability, nutrient limitation (N-limited vs. C-limited) and dilution rate (increasing or decreasing) we performed two enrichment tests for each TF, one for up-regulation of the target genes, and one for down-regulation of the target genes using adj. $p<0.05$ as cutoff and $logFC>0$ for up-regulation and $logFC<0$ for down-regulation. For the factors with more than two levels (carbon source and extra compound) we only performed one test for each TF.

Results

Controllability of transcriptional regulatory networks

The nine transcriptional regulatory networks (TRNs) that were included in this study are presented in Table 1, including three *S. cerevisiae* TRNs[9, 18], one *E.coli* network[19], two mouse networks[16] and three human networks[15, 16]. Each of these networks were analyzed in terms of network controllability[17] and stability. The networks were constructed using ChIP-chip and ChIP-seq datasets as a base and they contain only TF-TF interactions (i.e. non-TF genes were

filtered out). The number of driver nodes n_D is also presented in Table 1, with n_D being defined as the minimum number of input nodes in the network that need to be controlled in order to control 100% of the network[17]. As an example, by controlling node A in Figure 1a, we can control all three nodes, while we need to control node A, and C to have control over 100% of the TFs in Figure 1b. The example in Figure 1c contains an internal loop which means that 100% of the network can be controlled by controlling any of the nodes A, B or C as input. For a TRN the concept of network controllability corresponds to one or more TFs responding to environmental changes and these TF regulates other TFs by controlling their transcription. The number of driver nodes, n_D, was determined from the "maximal matching" graph (see materials and methods). In terms of stability, the system of control in Figure 1a and Figure 1b will always have asymptotically stable steady states, while the TF system in Figure 1c can have either unstable or stable steady states depending on the parameters and the type of regulation (activation or repression). For each network we studied how controllability and stability are related to its topology (random or scale-free) and its average degree, i.e. the average number of connections each TF has to other TFs in the network.

TRNs show circular control motifs

Error! Reference source not found. shows the number of driver nodes n_D as a measure of controllability of the nine different TRNs. For the Yeastract network only 17 % of the TFs need to be controlled in order to control all the other TFs. This large controllability is due to a large internal loop in the network containing 78% of the TFs where all the TFs in the loop are controlling each other in a circular fashion. We call this internal loop a circular control motif (CCM) and an example of a CCM can be found in Figure 1c where node A controls node B, node B controls node C and node C control node A in a circular manner. In terms of controllability this means that all the TFs in the loop can be controlled regardless which of the nodes is externally controlled. For the Harbison networks n_D is 37% of the TFs for the $p<0.005$ network and 48% for the $p<0.001$ network (these p-values indicate measure the statistical evidence of the interactions included in the network). We see the same trend in these networks; a CCM contains around 36% and 19% of the nodes in these two networks, respectively. For the *E.coli* network derived from Regulon DB, which has a n_D of 64%, we cannot find any large circular control network motif, and it is only possible to control a maximum of 6% of the nodes by controlling a single input node. For the mouse networks n_D is 4% and 5% for the Chea v.1 and Chea v.2 networks, respectively, whereas n_D is 24%, 8%

and 54% for the three Human networks. The number of nodes belonging to the CCM for the different networks is presented in Supplementary table 1.

Controllability of simulated networks

In order to test if the circular control motif (i.e. TFs controlling each other in a circular manner) in yeast, mouse and human, but not in *E.coli*, has evolved as a result of higher connectivity and higher average degree in these networks we simulated random Erdos-Renyi networks [22] with different average degree and scale-free Barabasi-Albert networks [21] with various average degree. The controllability for random Erdos-Renyi networks, when controlling one input node is dependent on the average degree of the network and for the simulated random networks with average degree 10 or higher we only need to control 1 TF as input (approx. 1%) to control 100% of the network (Figure 2). However, for the simulated scale-free networks the controllability is less dependent on the average degree of the network. The n_D for the scale-free network decreases slightly when increasing the average degree but does not exceed 44% for the simulated networks. In Figure 2 we have also plotted the number of driver nodes n_D as a function of the average network degree for the seven real networks included in this study. For the three *S.cerevisiae* networks the trend is that the networks with higher average degree also have lower n_D. The Yeastract network for example has an average degree of 18.28 and a n_D of 17%. The two mouse networks also have high average degrees and low n_D.

Stability analysis of transcriptional regulatory networks

To illustrate the stability of a TRN we will consider a simple network with 3 TFs. For each of these TFs we let n_i denote the number of copies of the i'th transcription factor in the network and r_i denote the rate of production of the TF. Assuming that the i'th TF does not regulate its own transcription the time evolution of n_i can be described by the equation $\frac{dn_i}{dt} = r_i - \delta_i n_i$

where δ_i is the specific degradation rate of the i'th TF. The rate r_i can be seen as a function of the abundances of all TFs that are regulating the transcription of the i'th TF in the regulatory network.

The differential matrix of the system can be expressed as $D = \frac{\partial \vec{r}}{\partial \vec{n}} - \Delta$ where $\partial \vec{r}/\partial \vec{n}$ is the Jacobean matrix for the transcription rates and Δ is a diagonal matrix with the specific degradation rates δ. To perform a stability analysis of the system we can calculate the

eigenvectors and eigenvalues of the differential matrix D, in the neighborhood of a steady state point. For the example for the network in Figure 1a the topological differential matrix will be:

$$D = \begin{pmatrix} -\delta_A & \frac{\partial r_B}{\partial n_A} & 0 \\ 0 & -\delta_B & \frac{\partial r_C}{\partial n_B} \\ 0 & 0 & -\delta_C \end{pmatrix}$$

Note that we have inserted zeros in the matrix for the instances where we do not have an edge in the regulatory network, e.g. TF C is not directly controlled by TF A so $\frac{\partial r_C}{\partial n_A} = 0$, etc. The Eigenvalues of this matrix is obtained by solving the equation det(D-λ)=0 and thus the eigenvalues for the network in Figure 1a becomes $\lambda_1 = -\delta_A,\ \lambda_2 = -\delta_B,\ \lambda_3 = -\delta_C$. If we assume that the degradation rates δ_A , δ_B and δ_C will be positive, all eigenvalues for the system will be negative, independent on the parameters and type of regulation, but clearly dependent only on the topological structure of the network. The fact that we will always have negative eigenvalues implicates that the system in example 1A will always be stable. Also for the example in figure 1B the eigenvalues will be $\lambda_1 = -\delta_A,\ \lambda_2 = -\delta_B,\ \lambda_3 = -\delta_C$, and this network is therefore also stable. For the example in Figure 1c the eigenvalues of the system will be a function of $\frac{\partial r_B}{\partial n_A}$, $\frac{\partial r_C}{\partial n_B}$ and $\frac{\partial r_A}{\partial n_C}$ and they can be either positive or negative depending on the parameters and the type of regulation. If all (real parts of the) eigenvalues are negative, the system will be stable, but if one or more of the real parts of the eigenvalues are positive, the system will be unstable. A more general derivation of this concept called topological stability analysis and another example with three transcription factors is presented in Electronic supplementary information (ESI) and Supplementary figure S1. This approach can be applied for a general network to calculate the supports of the eigenvectors (components with non-zero values) of the Jacobean matrix of the system, as a function only of the topology and regardless of the system parameters.

Hierarchical structure of the TRNs

The hierarchical structure of both the *E.coli* and *S.cerevisiae* transcriptional regulatory network has previously been reported[36, 37]. However, Figure 3a and b shows that the TRN for *E.coli* is more hierarchical than yeast when it comes to regulation. The *E.coli* regulation is controlled by *cRP* at the top of the hierarchical tree (Figure 3b), whereas for yeast there are

many TFs on the top hierarchical level (*SWI6, MBP1, FKH2, FKH1, ABF1, NRG1, INO4* and *SKN7*) (Figure 3a). Another difference between the *E.coli* and yeast network structure is the presence of the large circular control motif (CCM) in the yeast network where some of the TFs in this internal loop are even present in the top hierarchical level (*ABF1* and *MBP1*).

Figure 3c shows the human network (ChEA v.2) ordered in a hierarchical way, where four TFs are on top of the hierarchical structure (*TP63, SOX2, AR, GABP).* However, since 87% of the TFs in this network are part of the CCM this means that by controlling one of the top nodes, 87% of the network can be controlled. The TFs included in the CCM are marked in green in Figure 3c. Figure 3d shows the same hierarchical structure for the mouse (ChEA v.2) network and here is 92% of the TFs are part of the CCM.

Integrated analysis reveals condition-specific regulation in yeast

In order to identify TFs that respond to specific environmental cues that can represent input nodes in the network controllability analysis we analyzed transcriptome data from 233 yeast microarrays collected from 11 separate studies[24-35]. The experiments were all carried out in chemostat cultures where the environment could be tightly controlled, i.e. the specific growth rate, oxygen availability, carbon source etc. The data were analyzed using a regression model in order to describe the expression of each gene in terms of the effect of the environment.

The environmental factors included in the regression model are presented in **Error! Reference source not found.**. We were specifically interested in determination of the TFs that respond to these environmental factors. To identify condition-specific transcriptional regulation we considered three different features: i) TFs whose target genes are up- or down-regulated under different conditions, ii) TFs that do not change in expression between conditions and iii) TFs that are reported in literature to respond to a specific environmental cue.

For each environmental factor in **Error! Reference source not found.** we determined the genes where the environmental factor had a significant effect on the gene expression (adjusted p-value <0.05). We then performed a hypergeometric enrichment test for each TF in order to identify the TFs where the expression of the target genes was influenced by the environment. The target genes for each TF were defined by the TRN from Yeastract. The results for the

hypergeometric enrichment test for the factors oxygen availability, nutrient limitation and dilution rate and carbon source are presented in Supplementary Figure S2 and Table S2.

Figure 4 shows a comparison of feature i) and ii) above, i.e. the ability of the TF to change the expression of the target genes and the transcriptional change of the TF itself. The genes marked with red in the four plots are the TFs where the adj. p-value of the TF gene itself > 0.8 and the p-value of the hypergeometric test (transcriptional change of the target genes) < 0.05, i.e. TFs that don't change in expression themselves but their target genes are significantly regulated as a function of the environmental factor. These TFs must be regulated in another way as a response to the environment, either by activation through signaling or through interaction with other TFs. In response to oxygen availability these transcription factors are *HAA1* and *FKH2,* to nutrient limitation *MOT3*, and to altered specific growth rate (dilution rate) the TFs are *PHO4, FKH2* and *MGA1*, and in response to changes in carbon source they are *OAF1* and *CST6*.

We calculated the controllability of the network when controlling the TFs that were identified to respond to the environment in the TRN from Yeastract. When we control the nodes marked with red in Figure 4 for each environmental factor it is possible to control between 78-82% of the network, e.g. when controlling both *HAA1* and *FKH2* responding to oxygen limitation the controllability is 81%, 78% for nutrient limitation, 82% for dilution rate and 82% when controlling *OAF1* and *CST6* for the carbon source environmental factor.

Discussion

Regulation of cellular processes is complex and may occur on different levels in the cell. Here we consider transcriptional regulation, which can be thought of as consisting of three different layers. The first layer is the environment which can be thought of as different environmental cues, e.g. high or low oxygen levels, carbon or nitrogen limitation etc. The second layer consists of the transcription factors where some TFs respond to the environment and the TFs regulate the transcription of other TFs as well as other genes. The third layer consists of genes which encodes for proteins that carry out different functions in the cell. Here we analyzed the structure and topology of transcriptional regulatory networks (TRNs) by calculating the controllability, i.e. how many driver nodes are needed to control the networks.

Understanding the regulation of a cellular system on a global level can have large implications in for example metabolic engineering where an organism is engineered to

produce high amounts of a chemical of interest[38, 39]. If we have information about the regulation of metabolism this can help to identify bottlenecks in the route to over-production of the product of interest. Also, knowledge about the organization of transcriptional regulation will help in understanding of development of complex diseases and other complex traits.

The controllability analysis of the TRNs for *E.coli*, yeast, human and mouse resulted in several novel findings. The controllability of the *E. coli* TRN is low, i.e. we need to control as much as 64% of the input nodes to be able to control all output nodes, whereas the controllability for yeast, mouse and human is much higher. The organization of the *E.coli* TRN is hierarchical in the sense that most of the TFs do not control any other TFs but are controlled by one or more TFs and the number of TFs that are involved in one route of regulation from *cRP* on the top hierarchical level to the leaves of the tree is maximum 6 as can be seen in the *E.coli* hierarchical tree in Figure 3b. For the yeast, mouse and human networks we identify so called circular control motifs (CCMs) where we can control a large part of the network just by controlling one input node. The presence of these internal loops in the networks also means that the hierarchical organization of the network is less prominent since all nodes in the CCMs can be controlled just by controlling one of the nodes in the loop as input. For the human and mouse network around 90% of the network are interconnected and belongs to CCMs.

From analysis of the simulated networks presented in Figure 2 we can see that for a perfect scale-free network there are no CCMs, even for networks with a high average degree, since the corresponding network will follow a hierarchical tree where most of the nodes have a degree of 1 (i.e. controlled by one TF but not controlling any TFs) and there is one or few nodes that have a high degree (network hubs). The degree distribution of the scale free network follows a power law distribution, e.g. $P(k) \sim k^{-\alpha}$, where the parameter α typically has a value between 2 and 3. For the random network the degree distribution is uniformly distributed, i.e. the chance for having a node with a low degree is approximately equal to the chance of having a node with a high degree. For simulated random networks a high connectivity (i.e. average degree > 10) means that one can control 100% of the network through a single input node.

The behavior of the *E.coli* network seems to follow the behavior of the simulated scale-free network based on the controllability analysis and the results presented in Figure 2. We also investigated the degree distribution of the *E.coli* TRN and the network is scale-free in the

sense that it is possible to fit a power law distribution to the degree distribution (Supplementary figure S1). The mouse and human networks seem to be more random in the way the TFs are interconnected. Also, it is not possible to fit a power law distribution to these networks, which points to the fact that these networks are not scale-free. In the yeast networks the number of nodes with one or two neighbors is much less than expected for a scale-free network, especially for the TRN from Yeastract. The three different yeast networks included in this study differ in the confidence used to consider if a TF-gene interaction should be included in the network or not. The Harbison network uses either $p<0.001$ or $p<0.005$ for the binding probability cutoff from the ChIP-chip experiment. The Harbison $p<0.001$ network is the most conservative network and also has the lowest average degree. The Yeastract network is the least conservative yeast network and has the highest number of TF-gene interactions. It is not possible to compare TRNs from the different organisms (and even the different yeast TRNs), since the TRNs were constructed in different ways, with different confidence scores etc. However, still for all the different eukaryal TRNs there is a clear difference in structure and by using the different TRNs we could investigate the controllability of the network as a function of the average degree of the network, and hence also as a function of the confidence of TF binding scores, or how conservative the network construction process has been. The yeast network seems to behave in between the simulated scale-free and random networks, but as we increase the average degree (and become less conservative in what is considered as a TF binding event) the network behaves more random, and the maximum controllability when controlling one input node increases.

Network controllability has earlier been introduced by Liu et al. [17] where they also analyzed many different real networks in terms of controllability, including the yeast and *E.coli* TRNs. However the networks that were included in this study were different in the way they were constructed and the number of TF-gene interactions that were included. Here we use network controllability as a tool to study the topology and organization of the network and we identify CCMs in the yeast, mouse and human networks, but not in the *E.coli* network. In terms of evolution and the ability of the organism to adapt to changes in the environment it is interesting to see that the eukaryal TRNs included in this study seem to contain CCMs, but the bacterial TRN does not. Furthermore, based on our analysis it seems like more complex organisms have more random organization of their transcriptional regulatory network.

We find the presence of circular motifs in yeast, mouse and human interesting, as this may improve robustness in the transcriptional regulation when it comes to e.g. adapting to environmental changes.

In terms of stability a linear network motif will always be stable. A circular network motif, like the example presented in Figure 1c, on the other hand can be unstable, but also be more stable depending on the parameters and types of interaction. If two of the interactions in a triangular motif are positive (i.e. activation) and the third interaction is negative (i.e. repression) the steady state of the system will be less sensitive to changes in environmental parameters compared to a linear pathway. However, gene deletions or perturbations might disturb the system and make the steady state unstable, causing a rearrangement of the system similar to a non-equilibrium phase change, if an environmental or genetic parameter passes a certain threshold. A detailed discussion of these phenomena is contained in the supplementary material. According to our observations, more complex organisms contain more potentially unstable steady states of their TRN. This can contribute to a higher capability to maintain homeostasis under environmental changes, but also to sudden rearrangements of the state of the system as a result of mutations. Using the approach of topological stability analysis introduced in this paper we can identify the non-zero elements of the Jacobean matrix and find potentially unstable motifs in the TRN, without knowing the parameters of the system, but only as a function of the topology.

For the yeast TRN we identified TFs that respond to environmental cues by analyzing a large amount of microarray data from different controlled environments. The transcriptional regulation was found to be highly condition-specific and if we could identify condition-specific responses that allowed us to see how the yeast TRN looks under specific conditions or when changing environment from one state to another. This analysis shows that key TFs seem to exert a high degree of controllability in response to different environmental cues, i.e. a few TFs can control a large number of other TFs in the regulatory network.

In conclusion we perform analysis of the topology of TRNs for different species and find that there is an increasing complexity in terms of connectivity and controllability when moving from bacteria to yeast and further to mouse and human. Whereas the TRN for *E.coli* is scale-free the TRNs for eukaryotes seems to be more random, mainly due to the presence of circular control motifs (CCMs) involving a large number of TFs. These large CCMs enable control of a large fraction of the TRN through control of many single TFs, which may have been

important for establishing increased robustness towards different environmental changes. On the other hand the presence of CCMs can cause instability due to changes in binding affinity for the TFs in the motif, and this may result in large reprogramming of cellular function, e.g. resulting in disease development.

Acknowledgements

The authors thank Intawat Nookaew for valuable input regarding construction of the regulatory networks. Simulations were performed at Chalmers Centre for Computational Science and Engineering (C3SE). We acknowledge funding from the European Research Council project INSYSBIO (grant no. 247013), the Novo Nordisk Foundation and the Department of Energy (grant no. 5710003389).

Author's contributions

TÖ analyzed the data, constructed the networks, performed the controllability analysis and simulations and wrote the paper. SB derived the topological stability analysis. SB and JN supervised the work and edited the paper. All authors read and approved the manuscript.

References

1. P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee and E. M. Marcotte, *Current opinion in structural biology*, 2004, 14, 292-299.
2. I. Thiele and B. Ø. Palsson, *Nature protocols*, 2010, 5, 93-121.
3. T. Österlund, I. Nookaew and J. Nielsen, *Biotechnol Adv*, 2012, 30, 979-988.
4. D. McCloskey, B. Ø. Palsson and A. M. Feist, *Molecular systems biology*, 2013, 9.
5. M. W. Covert and B. Ø. Palsson, *Journal of Biological Chemistry*, 2002, 277, 28058-28064.
6. M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard and B. O. Palsson, *Nature*, 2004, 429, 92-96.
7. M. Herrgård, B. Lee, V. Portnoy and B. Palsson, *Genome research*, 2006, 16, 627.
8. G. Chua, M. D. Robinson, Q. Morris and T. R. Hughes, *Current opinion in microbiology*, 2004, 7, 638-646.
9. 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 and J. Yoo, *Nature*, 2004, 431, 99-104.
10. M. Vidal, M. E. Cusick and A.-L. Barabasi, *Cell*, 2011, 144, 986-998.
11. 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 and I. Simon, *Science*, 2002, 298, 799-804.
12. T. R. Hughes and C. G. de Boer, *Genetics*, 2013, 195, 9-36.
13. S. Chandrasekaran and N. D. Price, *Proceedings of the National Academy of Sciences*, 2010, 107, 17845-17850.
14. I. Dunham, E. Birney, B. R. Lajoie, A. Sanyal, X. Dong, M. Greven, X. Lin, J. Wang, T. W. Whitfield and J. Zhuang, 2012.
15. F. Zambelli, G. M. Prazzoli, G. Pesole and G. Pavesi, *Nucleic acids research*, 2012, 40, W510-W515.

16. A. Lachmann, H. Xu, J. Krishnan, S. I. Berger, A. R. Mazloom and A. Ma'ayan, *Bioinformatics*, 2010, 26, 2438-2444.
17. Y.-Y. Liu, J.-J. Slotine and A.-L. Barabási, *Nature*, 2011, 473, 167-173.
18. M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira and I. Sá-Correia, *Nucleic acids research*, 2006, 34, D446-D451.
19. S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo and A. López-Fuentes, *Nucleic acids research*, 2011, 39, D98-D105.
20. G. Csardi and T. Nepusz, *InterJournal, Complex Systems*, 2006, 1695.
21. A.-L. Barabási and R. Albert, *science*, 1999, 286, 509-512.
22. P. Erdős and A. Renyi, *Acta Mathematica Hungarica*, 1961, 12, 261-267.
23. R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew and J. Nielsen, *PLoS computational biology*, 2013, DOI: 10.1371/journal.pcbi.1002980.
24. T. Knijnenburg, J.-M. Daran, M. van den Broek, P. Daran-Lapujade, J. de Winde, J. Pronk, M. Reinders and L. Wessels, *BMC genomics*, 2009, 10, 53.
25. C. Cipollina, J. van den Brink, P. Daran-Lapujade, J. T. Pronk, M. Vai and J. H. de Winde, *Microbiology*, 2008, 154, 337-346.
26. D. A. Abbott, T. A. Knijnenburg, L. M. de Poorter, M. J. Reinders, J. T. Pronk and A. J. van Maris, *FEMS yeast research*, 2007, 7, 819-833.
27. J. Aguilera, T. Petit, J. H. de Winde and J. T. Pronk, *FEMS yeast research*, 2005, 5, 579-593.
28. R. De Nicola, L. A. Hazelwood, E. A. De Hulster, M. C. Walsh, T. A. Knijnenburg, M. J. Reinders, G. M. Walker, J. T. Pronk, J. M. Daran and P. Daran-Lapujade, *Appl Environ Microbiol*, 2007, 73, 7680-7692.
29. V. M. Boer, S. L. Tai, Z. Vuralhan, Y. Arifin, M. C. Walsh, M. D. Piper, J. H. de Winde, J. T. Pronk and J. M. Daran, *FEMS yeast research*, 2007, 7, 604-620.
30. T. A. Knijnenburg, J. H. de Winde, J. M. Daran, P. Daran-Lapujade, J. T. Pronk, M. J. Reinders and L. F. Wessels, *BMC Genomics*, 2007, 8, 25.
31. S. L. Tai, V. M. Boer, P. Daran-Lapujade, M. C. Walsh, J. H. de Winde, J. M. Daran and J. T. Pronk, *Journal of Biological Chemistry*, 2005, 280, 437-447.
32. P. Daran-Lapujade, M. L. Jansen, J. M. Daran, W. van Gulik, J. H. de Winde and J. T. Pronk, *The Journal of biological chemistry*, 2004, 279, 9125-9138.
33. P. Daran-Lapujade, S. Rossell, W. M. van Gulik, M. A. Luttik, M. J. de Groot, M. Slijper, A. J. Heck, J. M. Daran, J. H. de Winde, H. V. Westerhoff, J. T. Pronk and B. M. Bakker, *Proc Natl Acad Sci U S A*, 2007, 104, 15753-15758.
34. A. Fazio, M. Jewett, P. Daran-Lapujade, R. Mustacchi, R. Usaite, J. Pronk, C. Workman and J. Nielsen, *BMC genomics*, 2008, 9, 341.
35. B. Regenberg, T. Grotkjær, O. Winther, A. Fausbøll, M. Åkesson, C. Bro, L. K. Hansen, S. Brunak and J. Nielsen, *Genome biology*, 2006, 7, R107.
36. H. Yu and M. Gerstein, *Proceedings of the National Academy of Sciences*, 2006, 103, 14724-14731.
37. H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer and A.-P. Zeng, *Nucleic acids research*, 2004, 32, 6643-6649.
38. E. J. Kerkhoven, P. J. Lahtvee and J. Nielsen, *FEMS yeast research*, 2014.
39. M. Li and I. Borodina, *FEMS yeast research*, 2014.

Figure legends

Figure 1 – Network controllability and stability. Concept of network controllability. (A) To control all three nodes, it is enough to control node A in the left network (n_D=1). (B) two driver nodes are needed to control the middle network (n_D=2). (C) Since node A, B and C are connected in a circular control motif (CCM) we can choose any of the three nodes as driver node and still have 100% control over the network. In terms of stability network (A) and (B) will always be stable but network (C) can possibly be unstable.

Figure 2 – Controllability when controlling one input node for simulated and real networks. The x-axis shows the average degree of the network and the y-axis shows the controllability (relative amount of driver nodes needed to control 100% of the network).

Figure 3 – Hierarchical structure of the transcriptional regulatory networks. (A) *S.cerevisiae* Harbison network. (B) *E.coli* RegulonDB network (C) Human ChEA v.2 network (D) Mouse ChEA v.2 network

Figure 4 – TFs responding to environment. The x-axis in each plot shows the hyper-geometric p-value for each TF based on the expression of the target genes. A low hyper-geometric p-value indicates that the environmental factor has an effect on the expression of the target genes. The y-axis in each plot shows the adjusted p-value of the TF gene, a high p-value indicates that the gene is not changed in expression as a function of the environmental change. The TFs marked with red have a hyper-geometric p-value for the target genes less than 0.05 and a adjusted p-value for the TF gene greater than 0.8 (A) Oxygen availability. (B) Nutrient limitation. (C) Dilution rate. (D) Carbon source.