Evolutionary systems biology: methods to reconstruct and compare transcriptional regulatory networks

M. Madan Babu[1*], Benjamin Lang[1] & L. Aravind[2]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB22QH, U.K.
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, U.S.A

***For correspondence:**
Email: madanm@mrc-lmb.cam.ac.uk
Phone: 44-1223-402208
Fax: 44-1223-213556

Length: 143 (abstract) + 4092 (main text + figure legend)
Figures: 7

Abstract

The availability of entire genome sequences and the wealth of literature on gene regulation have enabled researchers to model an organism's transcriptional regulation system in the form of a network. In order to reconstruct such networks in non-model organisms, three principal approaches have been taken. Firstly, one can transfer interactions between homologous components from a model organism to the organism of interest. Secondly, microarray experiments can be used to detect patterns in gene expression that stem from regulatory interactions. As a third approach, knowledge of experimentally characterized transcription factor binding sites can be used to analyze the promoter sequences in a genome in order to identify potential binding sites. In this chapter, we will focus in detail on the first approach and describe methods to reconstruct and analyze transcriptional regulatory networks of uncharacterized organisms by using a known regulatory network as the template.

Keywords: transcriptional regulatory network; network reconstruction; template based method; network motif; life style; statistical significance

1.0 Introduction

Advancements in genome sequencing techniques are yielding the complete sequences of genomes of several organisms. Although the methods to predict functional coding genes within these genomes are highly advanced today, what this information does not tell us, however, is how the products of these genes interact and how they are regulated. In recent years, a variety of high-throughput techniques have been developed and employed to generate vast amounts of data that could be used to bridge this gap. For instance, high-throughput microarray experiments are providing expression data for a number of genes under a variety of conditions (1-3); large-scale experiments using chromatin immuno-precipitation combined with microarray hybridization (ChIP-chip) are giving us specific evidence of regulatory proteins binding to stretches of DNA (4-7) and additionally, large amounts of data on protein-DNA interactions from individually conducted experiments over the years are collected in databases, resulting in a wealth of literature on gene regulation (8, 9). For some model organisms, such as *E. coli* and yeast, these data have been integrated to produce comprehensive models of their transcriptional regulatory interactions in the form of networks (10, 11). The challenge that we currently face is to develop computational techniques that would allow us to make the most of this information to understand regulation in organisms that are less well characterized.

There are three fundamental approaches that can be taken to infer the structure of the organism's regulatory interaction network from these data, at varying levels of resolution. These are (i) **Template based methods:** This approach exploits the principle that orthologous transcription factors generally regulate the expression of orthologous target genes. Thus, in this network reconstruction method, one starts with a known regulatory network and transfers information about interactions to genes that have been determined to be orthologous in a target genome of interest (12-14). (ii) **Reverse engineering using gene expression data:** In this approach, one scans for patterns in gene expression data from time-series experiments and from experiments conducted across several different conditions. If a gene is upregulated following an increased production of a transcription factor, or down-regulated following a knockout of a transcription factor, a regulatory interaction between the two is inferred. In the case of expression analysis over different experimental conditions, one infers sets of genes with a similar expression profile across many conditions to be co-regulated by the same set of transcription factors (15-19). Such inferences become more accurate as the number of measurements over a certain period of time (the time-scale resolution of the data) increases since this allows direct regulatory interactions to be distinguished from indirect (multi-step) regulation. (iii) **Inferring networks by predicting cis-regulatory elements:** The third approach makes use of information about experimentally well characterized transcription factor binding sites to make inferences about regulatory interactions. In this method, promoter regions in the genome of interest are scanned using known binding site profiles of characterized transcription factors. The set of genes which are predicted to have a binding site are hypothesized to be regulated by the corresponding transcription factor (20-23).

While the methods mentioned above exploit three different principles, there have been considerable efforts to develop a combined approach to predict regulatory interactions with a higher degree of confidence (10, 24-28). For instance, while analyzing microarray expression data, the initially determined sets of co-regulated genes can be refined by investigating whether or not the same transcription factor actually binds to all of them by predicting presence or absence of a binding site in the promoter regions of these genes. In this way, we can distinguish

directly regulated genes from ones that are regulated through more complicated network motifs or even genes that just randomly happen to show a similar expression profile.

In this chapter, we will primarily focus in detail on the template-based method (12). In order to know more about the other methods discussed, the reader is asked to refer to the other chapters in this book which explicitly deals with network reconstruction procedures using gene expression data and binding site data. Alternatively, the reader is suggested to visit http://www.mrc-lmb.cam.ac.uk/genomes/madanm/blang/methods/. In this supplementary material website, we have provided a comprehensive review of published work that exploits these different methods.

2.0 Materials

2.1 Hardware requirements

1. Personal Computer with at least 512MB memory, 10GB hard-disk space, and a processor of at least 1GHz or better.
2. Access to a Linux or a UNIX workstation
3. Stable connection to the internet

2.2 Software requirements

1. A recent version of PERL installed on a Linux or a UNIX environment (freely available for download at: http://www.perl.com)
2. A versatile Windows text editor such as TextPad (freely available at: http://www.textpad.com)
3. A recent version of NCBI BLAST suite of programs (29) for Linux (freely available from the NCBI website at: http://www.ncbi.nlm.nih.gov/Ftp/)
4. A recent version of the motif finding program MFINDER (30) for windows (freely available from the Weizmann Institute at: http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html)

2.3 Template transcriptional regulatory network

Information on transcriptional regulation is available for several model organisms. The following websites provide a list of regulatory interactions in the different organisms. The networks have been manually curated in several cases and contain regulatory interactions inferred from large-scale functional genomics experiments in the case of yeast. For the method described in this chapter, we only use only the *E. coli* regulatory network as the template. It should be noted at this point that any network can be potentially used as a template network.

1. *Escherichia coli*: RegulonDB (8) (http://regulondb.ccg.unam.mx/index.html)
2. *Bacillus subtilis*: DBTBS (31) (http://dbtbs.hgc.jp/)
3. *Corynebacterium* species: Coryneregnet (9) (https://www.cebitec.uni-bielefeld.de/groups/gi/software/coryneregnet/)
4. *Saccharomyces cerevisiae*: A curation of regulatory interactions from several different small-scale and large-scale studies (32, 33) (http://www.mrc-lmb.cam.ac.uk/genomes/madanm/tfcomb/tnet.txt)

2.4 Complete genome sequences and lifestyle information for organisms of interest

1. The complete genome sequence and the predicted proteome of several prokaryotic and eukaryotic genomes can be obtained from the NCBI genomes website at: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

2. Detailed and systematic information about the lifestyle of the different completely sequenced genomes can be obtained from the same website by clicking on the “organism info” tab. If this information is not available, the reader is suggested to obtain it from the publication describing the genome sequence or refer to the Brock's Biology of Microorganisms or Bergey’s Manual of Determinative and Systematic Bacteriology (available at: http://www.cme.msu.edu/Bergeys/)

3.0 Methods

The set of all transcriptional regulatory interactions within a cell can be conceptualized as a graph which is best modeled as a network (10, 11). In such a network, nodes represent genes that are transcription factors or targets and edges represent direct transcriptional regulatory interaction. A number of recent studies on transcriptional networks in prokaryotes and eukaryotes have shown that the structure of such networks can be differentiated into three distinct levels of organization (10). At the most basic level, the network consists of a single regulatory interaction between a transcription factor and its target gene (Fig 1a). At the intermediate level of organization, studies have uncovered that the basic unit is organized into fundamental building blocks of transcriptional regulation, called network motifs (Fig 1b). These motifs are discrete functional units and are defined as small patterns of interconnections that are seen in several different contexts within the network (6, 34). Finally, at the global level of organization, the set of all transcriptional regulatory interactions in a cell form the global structure and has been shown to have a hierarchical and a scale-free topology (35-37). In other words, such a global structure is characterized by the presence of many transcription factors which regulate few genes and the presence of a few transcription factors, the regulatory hubs, which regulate many genes (Fig 1c).

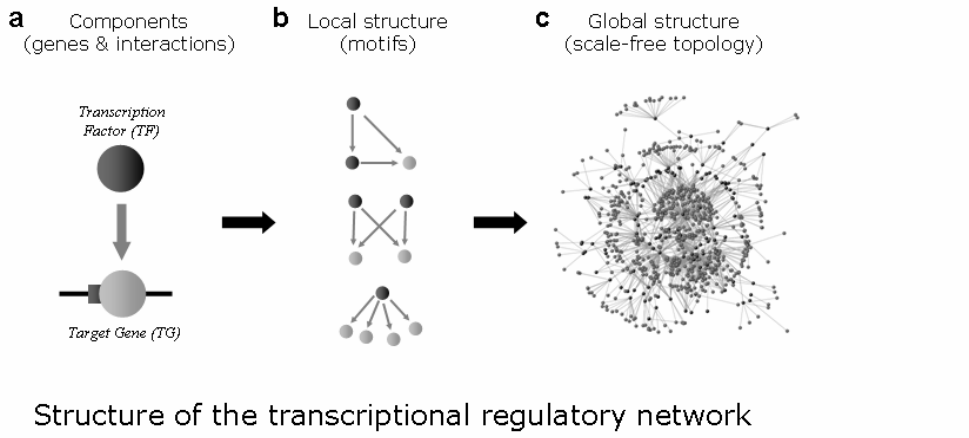

Figure 1: Organization of the transcriptional regulatory network into (a) components, (b) local structure and (c) global structure. Black and gray circles represent transcription factors (TFs) and target genes (TGs) and an arrow represents a direct regulatory interaction.

In the following section, we describe the network reconstruction procedure to reconstruct conserved transcriptional regulatory interactions in a genome of interest using a template network (12). Having obtained the reconstructed networks, we then describe methods to analyze the networks at different levels of organization and methods to assess significance of the evolutionary conservation. Finally, we will also describe methods to correlate the conserved network structure with the lifestyle of the organism in order to obtain insights into interactions that are particularly

important for the organism of interest. Throughout this section, we will be describing the methods by using the *E. coli* transcriptional network as the template network.

3.1 Procedure to reconstruct the transcriptional network in a genome of interest using a template network

3.1.1 Network reconstruction procedure

1. The transcriptional regulatory network for *E. coli* was used as the basis to reconstruct networks for other genomes. Information about regulatory interactions was obtained from RegulonDB (8). Thus the template network consisted of 1295 transcriptional interactions involving 755 proteins (112 transcription factors).
2. Orthologous proteins were identified in the genome of interest using the method described below. If orthologs were identified for an interacting transcription factor and target gene, then an interaction was inferred to be present in the genome of interest (Fig 2). Note that this method can be readily extended to any starting template network.

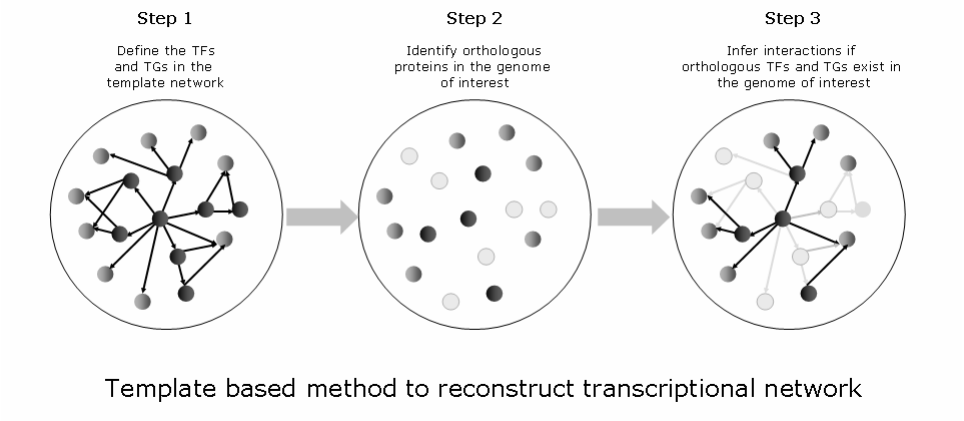

Figure 2: Method to reconstruct the transcriptional network for a genome of interest starting from an experimentally characterized template regulatory network. Black circles represent transcription factor proteins in the network, gray circles represent target genes and black arrows represent direct regulatory interaction. Light gray circles represent proteins that are absent in the genome of interest for which an interaction is known in the template network.

3.1.2 Orthology detection procedure

Detecting orthology is a non-trivial exercise and can be confounded by paralogs or sequence divergence in a genome of interest. After testing various orthology detection procedures (bi-directional best hit, best hits with defined e-value cut-offs, etc), we arrived at a hybrid procedure that was used to reliably identify orthologous proteins in a genome of interest (see note 1 and Fig 3).

3.1.2.1 Bi-directional best-hit procedure

1. For each protein **P** in the template network, a BLAST search was performed against the genome of interest (**x**).
2. The best hit, sequence $\mathbf{P^x}$ from genome **x**, was then used as a query and a BLAST search was carried out against the *E. coli* genome.
3. If the best hit using $\mathbf{P^x}$ as the query happens to be **P** in the template genome, then **P** & $\mathbf{P^x}$ were considered as orthologous proteins.
4. If however $\mathbf{P^x}$ does not pick up **P** from the template genome as its best hit, then a BLASTclust procedure was adopted.

3.1.2.2 BLASTclust procedure

1. For each of the proteins **P** in template network for which the above-mentioned procedure did not pick up orthologous proteins, the best-hit sequences $\mathbf{P^x}$ (using **P** as the query against genome **x**) for each genome was obtained. Thus, for every protein **P**, this procedure gave us a set of proteins, which were the best hits from genomes where the bi-directional best-hit procedure failed.
2. The set of sequences thus obtained along with the query protein was taken through a BLASTclust (38) procedure using length conservation (L) of 60% and a score density (S) of 60% (See notes 2-4).
3. All the sequences belonging to the cluster that also contains the query protein **P** from the template network were considered as orthologs (Fig 3).

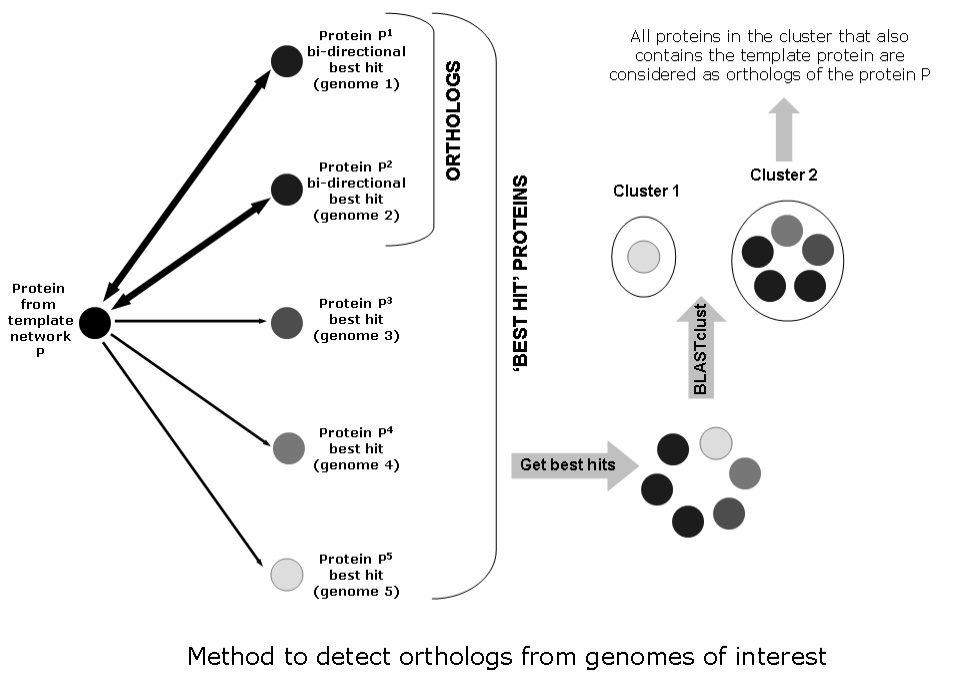

Figure 3: Hybrid method to detect orthologous proteins from genomes of interest

3.1.3 Method to create random networks to assess significance of trends observed in reconstructed networks

To assess the significance of the trends observed in the real network, it is essential to ensure that the observed trends are meaningful and are not something that is expected by chance. To this

end, generation of random networks provides a good way of assessing the statistical significance of the trend. The method below (fig 4) details the procedure to generate several random networks similar to what is seen in the reconstructed network. The random networks generated will be used in the next sections to explain how statistical significance is computed.

1. The number of transcription factors and target genes were defined in the reference network
2. Orthologs of the transcription factors and target genes were detected in the genome of interest and the numbers of TFs, x and TGs, y are noted.
3. To generate the random network, x TFs and y TGs were randomly chosen and the network was reconstructed based on the randomly chosen x TFs and y TGs.

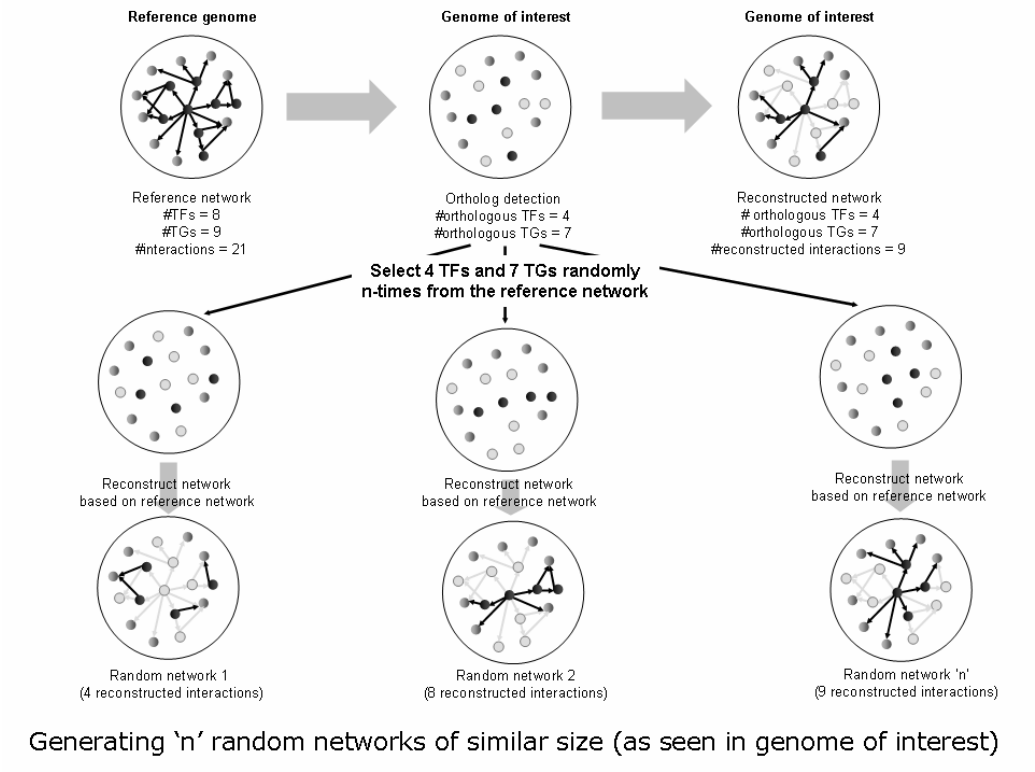

Figure 4: Procedure to generate random networks to assess statistical significance

3.2 Methods to analyze genes and regulatory interactions

3.2.1 Procedure to analyze conservation of genes and regulatory interactions

1. Interactions in the template network were ordered and indexed. For every genome, the reconstructed network was represented as a vector of 1 and 0. 1 represents the presence of an interaction and 0 represents the absence of an interaction. Note that a similar vector can be generated to create a transcription factor presence/absence profile for all the genomes of interest.
2. Having constructed the vector for every genome, the distance between the vectors (which represents the similarity in the interactions conserved between a pair of genomes) was calculated

(Fig 5). A tree representing the similarity of interactions (or genes) conserved in the different genomes was obtained using the distance matrix.
3. Alternatively, the vectors were be clustered using standard clustering programs such as cluster (39) and were visualized using matrix2png (40). This provided a visual representation of the interactions conserved in genomes of interest (see notes 5-7).
4. Note that a similar exercise performed on the presence/absence profile for transcription factors would allow us to group genomes based on similar transcription factor content.

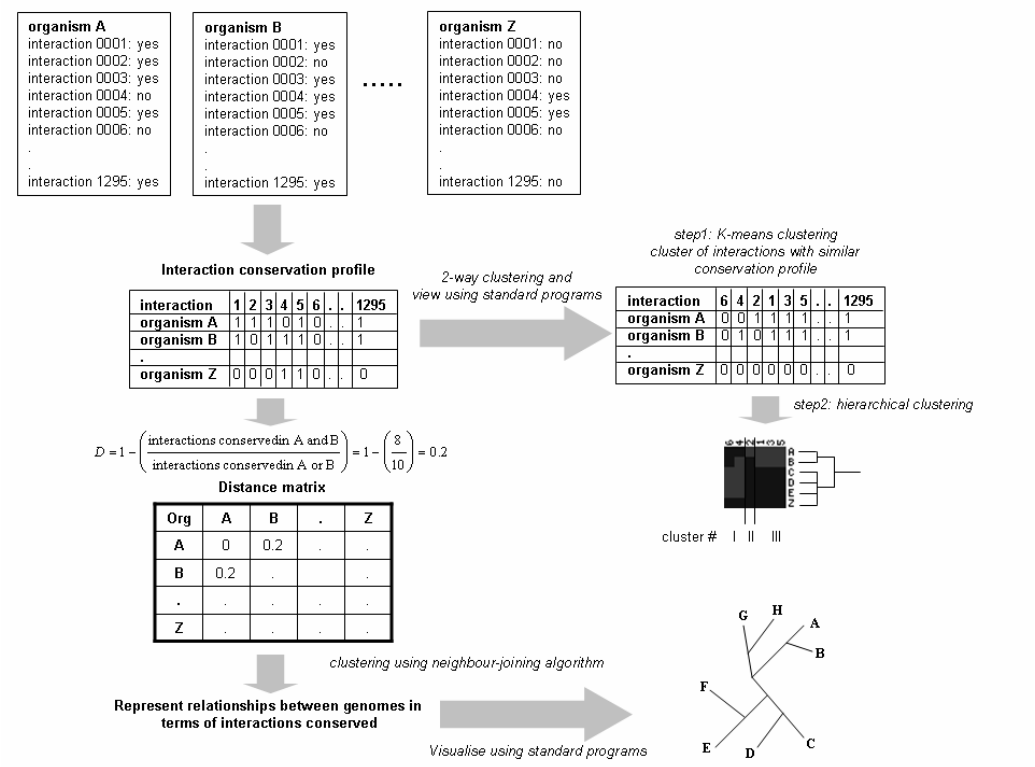

Figure 5: Method to analyze interactions and genes conserved in genomes of interest

3.2.2 Procedure to analyze significance of conservation of genes and interactions

To assess if a genome of an organism living in a particular environment, has conserved regulatory interactions differently from what would be expected by chance, we employ the following procedure.
1. For each of the genome of interest, we generated 10,000 random networks as described in section 3.1.3
3. For each of the genomes of interest, the mean μ and standard deviation σ of the fraction of interactions (or genes) conserved for the 10,000 random networks and the reconstructed network were obtained.
4. The P-value, a measure of statistical significance, was calculated as the fraction of the runs where the fractional conservation was greater than or equal to the observed value for a genome of interest.
5. The Z-score, a measure of how significantly the value deviates from the expected value, was calculated as $Z = (\mu^{obs} - \mu^{mean})/\tilde{\sigma}$

3.3 Methods to analyze local network structure

Studies on the local level of transcriptional regulatory network have elucidated the presence of small patterns of interconnections called network motifs. It is now generally accepted that three kinds of network motifs dominate these networks in prokaryotes and eukaryotes. These are the (i) feed forward motif, (ii) single input motif and (iii) multiple input motif. In the following section, we describe the procedure to analyze the conservation of network motifs in the genomes of interest.

3.3.1 Procedure to analyze conservation of network motifs

1. Network motifs in the template network were identified using the Mfinder program.
2. All identified motifs in the template network were ordered and indexed
3. A motif was considered to be absolutely conserved in a genome, if all the genes constituting the motif in the template network were conserved in the genome of interest. If some genes were missing, the fraction of conserved interactions in the motif was noted.
4. Thus for each genome, an ordered n-dimensional vector (motif conservation profile) is created, where n is the number of motifs considered. The values represent the fraction of the interactions forming the motifs that are conserved.
5. This matrix was then subjected to the procedure explained in section 3.2.1 to identify organisms that have a similar motif conservation profile.

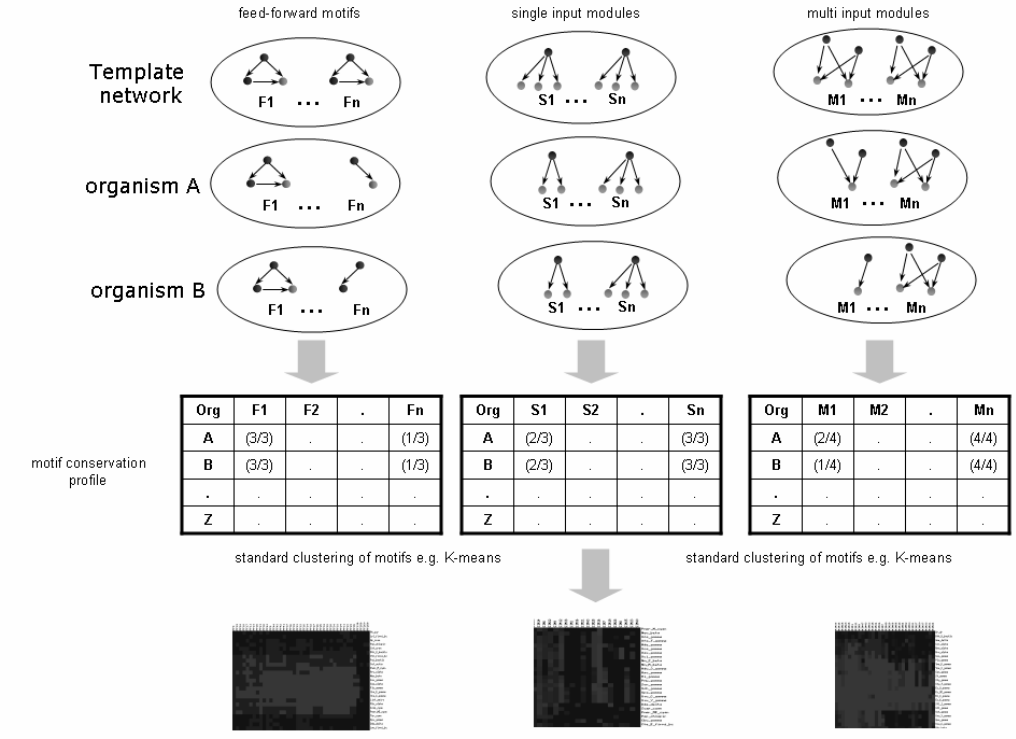

Figure 6: Procedure to analyze conservation of network motifs in genomes of interest

3.3.2 Procedure to analyze significance of conservation of network motifs

1. In order to assess if interactions in motifs are selectively conserved, it becomes important to evaluate whether interactions in a motif are more conserved than any interaction in the network. We introduce a term called conservation index (C.I.) that allows us to assess this trend.

$$C.I._{genome\ X} = \log_2\left(\frac{\mathrm{R}_{\mathrm{motif}}}{\mathrm{R}_{\mathrm{all}}}\right)$$

$$\mathrm{R}_{\mathrm{motif}} = \frac{\mathrm{I}^{\mathrm{motif}}_{genome\ X}}{\mathrm{I}^{\mathrm{motif}}_{template}} \quad \text{and} \quad \mathrm{R}_{\mathrm{all}} = \frac{\mathrm{I}^{\mathrm{all}}_{genome\ X}}{\mathrm{I}^{\mathrm{all}}_{template}}$$

In this definition, $\mathrm{I}^{\mathrm{motif}}_{genome\ X}$ is the number of interactions that forms a motif in the template network, which has been conserved in genome X. $\mathrm{I}^{\mathrm{motif}}_{template}$ is the number of interactions in a motif in the template network. $\mathrm{I}^{\mathrm{all}}_{genome\ X}$ is the total number of interactions that have been conserved in genome X and $\mathrm{I}^{\mathrm{all}}_{template}$ is the total number of interactions in the template network (see notes 8).

2. To assess if the C.I. value could be obtained by chance, the same value was calculated for 10,000 random networks generated using the procedure described in section 3.1.3.

3. For each of the genomes of interest, the mean μ and standard deviation σ of the C.I. value for the 10,000 random networks were obtained.

4. The P-value, a measure of statistical significance, was calculated as the fraction of the runs where the value of C.I. was greater than or equal to the observed value for a genome of interest.

5. The Z-score, a measure of how significantly the value deviates from the expected value, was calculated as $Z = (\mu^{obs} - \mu^{mean})/\tilde{\sigma}$

3.4 Methods to analyze global network structure

3.4.1 Procedure to analyze global network structure

The distribution of outgoing connectivity provides an indication about the large-scale structure of networks. It is well established that the outgoing connectivity for the *E. coli* network follows a scale-free behaviour, i.e. the distribution is best approximated by a power-law function $\mathbf{T} = a\mathbf{K}^{-b}$ where T is the number of transcription factors with **K** connections. To evaluate the distribution for the reconstructed networks we describe the following procedure:

1. For each of the reconstructed networks of the genomes of interest, the distribution was approximated by a linear function ($\mathbf{T} = a + b\mathbf{K}$), exponential function ($\mathbf{T} = ae^{-\mathbf{K}b}$; $\log \mathbf{T} = \log a - \mathbf{K}b \log e$) and a power-law function ($\mathbf{T} = a\mathbf{K}^{-b}$; $\log \mathbf{T} = \log a - b \log \mathbf{K}$).
2. The function that best approximates the observed distribution is identified was the one that has the lowest standard error.

3.4.2 Procedure to analyze significance of the conservation of global network structure

To identify the trend in random networks the following procedure was carried out:

1. For each of the genomes of interest, we a created 10,000 random networks as described in 3.1.3.
2. The procedure explained above (3.4.1) was executed on each of the 10,000 networks to get the function that best approximates the distribution for the random networks.
3. For each of the genomes of interest, the mean μ and standard deviation σ for the power-law exponent over all 10,000 random networks were computed.
4. As before, the P-value was calculated as the fraction of the runs where the value for the exponent was greater than or equal to the observed value.
5. As before, the Z-score was calculated as $Z = (\mu^{obs} - \mu^{mean})/\tilde{\sigma}$

3.5 Method to co-relate lifestyle data with conservation of regulatory interactions and network motifs

3.5.1 Life style based network similarity index (LSI)

1. For each organism studied, lifestyle information was collected from the literature and from various sources, including the NCBI genome information website, Brocks Manual of Microbiology and Bergey's Manual of Determinative and Systematic Bacteriology.
2. The following attributes were used to define the lifestyle class of an organism (see note 9)
    a. Oxygen requirement (aerobic, anaerobic, facultative, microaerophilic)
    b. Optimal growth temperature (hyperthermophilic, thermophilic, mesophilic)
    c. Environmental condition (aquatic, host-associated, multiple, specialized, terrestrial)
    d. Pathogen (yes, no)
3. The lifestyle (LS) of an organism is defined as a combination of the above four properties. For example, *E. coli* would be classified as "facultative:mesophilic:host-associated:no".
4. Similarity measure between any two organisms was defined as the similarity in the "interaction conservation profile" or the "network motif conservation profile".
5. Normalized similarity based on the interaction and motif conservation is calculated for each pair of life-style classes (Fig 7):

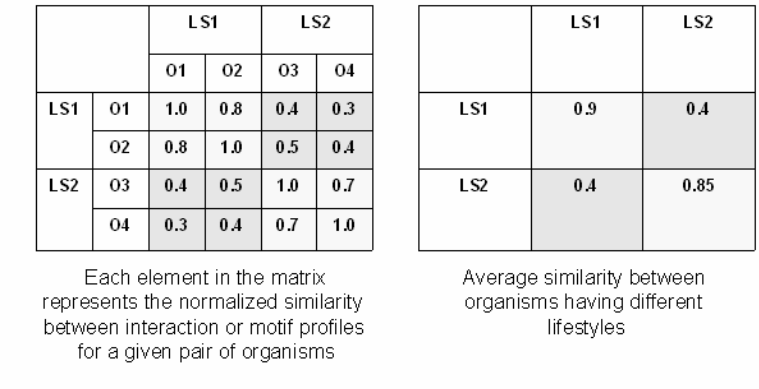

Figure 7: Life style and similarity in interactions conserved. LS: lifestyle class, O: organism.

6. Lifestyle based network similarity index (LSI), which is an indication of how often organisms with similar lifestyle conserve similar interactions or network motifs (see notes 9), was calculated as:

$$\mathrm{LSI} = \frac{\text{Average similarity between organisms belonging to the same lifestyle}}{\text{Average similarity between organisms belonging to different lifestyles}} = \frac{\dfrac{\sum \text{Diagonal elements}}{\text{Number of diagonal elements}}}{\dfrac{\sum \text{Off diagonal elements}}{\text{Number of off diagonal elements}}}$$

3.5.2 Procedure to assess significance of the observed LSI values

To test the significance of the LSI values, we perform randomization experiments.
1. First, 176 random networks of the size similar to each genome studied were generated.
2. Next, the LSI value was calculated for the random network using the definition of the lifestyle class defined in section 3.5.1.
3. This procedure was carried out 1000 times, and the *p*-value was calculated as the number of times the LSI values in the simulation were greater than the observed value.
4. The Z-score was calculated as the ratio of the difference between the observed and the average LSI value to the standard deviation in the observed distribution of LSI values for the 10,000 random networks.

4.0 Notes

1. Bi-directional best hit is a very conservative approach to detect orthologs. It performs best for closely related organisms and may fail to pick up orthologs from distantly related organisms. The best hit method using specific cut-offs is too liberal, and may result in false positive hits when the genomes compared are distantly related or when there are many closely related paralogs in the genome of interest. So our hybrid orthology detection method uses a combination of both methods as described above.

2. The BLASTclust procedure first carries out an all-against-all sequence comparison and produces clusters of sequences using the single linkage-clustering algorithm. This will ensure that orthologous proteins in distantly related organisms will still be picked up reliably through the sequences from the intermediately distant genomes.

3. Manual analysis of the clusters using various combinations of values for the parameters reveal that the parameters score density, **S**=0.6 and length overlap, **L**=0.6 performs best with an optimum coverage and lowest false positive rate.

4. In the BLASTclust algorithm, score density (**S**) is defined as the ratio of the number of identical residues in the alignment to the length of the alignment. Detailed documentation for BLASTclust is available at: ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastclust.html

5. Cluster is a wonderful program that allows users to cluster vectors representing the conserved network. It provides several clustering methods such as (i) hierarchical clustering (ii) K-means clustering and (iii) self organizing maps. Hierarchical clustering can be done using (a) single linkage (b) multiple linkage (c) centroid linkage and (iv) average linkage methods. The program also allows the use of different distance measures to cluster vectors such as Pearson's correlation coefficient, Euclidean distance, Spearman's rank correlation, etc. In our experience, we find that either K-means clustering or hierarchical clustering using the single linkage method

and Pearson's correlation coefficient distance measure gives the best results. The cluster software can be downloaded from: http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

6. Matrix2png is a simple and powerful program that can be used to visualize the vector representation of the conserved networks in the genomes of interest. It generates PNG format images from tab-delimited text files of vector data. This software can be downloaded from: http://www.bioinformatics.ubc.ca/matrix2png/

7. The distance matrix representing the similarity between the vectors representing the conserved networks can be visualized as a tree by using the treeview package. Treeview can be downloaded from: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

8. In the calculation of the conservation index (C.I.) for network motifs, the $\log_2$ of the ratio ensures that selection for and against are represented symmetrically in the graph. For example, if $R_{motif} = 0.9$ and $R_{all} = 0.6$, C.I. can be calculated as $\log_2 (.9/.6) = 0.58$. Thus if interactions in motifs are selected for, then the C.I value will be greater than 0, if not, the value will be less than 0.

9. It should be noted that other features such as salinity, pressure, tolerance to damaging radiation can also be used as additional attributes to define lifestyle class. It is worth mentioning our observation that organisms with similar lifestyle can be very distantly related and organisms that are close evolutionary relatives very often tend to colonize different ecological niches.

10. In the example described in section 3.5.1, LSI can be calculated as: [(0.90 + 0.85)/2] / [(0.4 + 0.4)/2] = 2.18 (i.e. the ration of the average of the diagonal elements to the average of the off-diagonal elements). In other words, if the organisms with a similar life style have higher similarity in motif or interaction content than organisms with dissimilar life style, then the LSI should be greater than 1.

Acknowledgements

MMB acknowledges the Medical Research Council, UK for financial support. LA acknowledges the Intramural Research Program of the NIH, NLM, NCBI, USA for funding. We thank Arthur Wuster for critically reading this manuscript.

Supplementary information

See http://www.mrc-lmb.cam.ac.uk/genomes/madanm/blang/methods/ for an overview of literature in the field of transcriptional network reconstruction.
See http://www.mrc-lmb.cam.ac.uk/genomes/madanm/evdy/ for detailed supplementary material and reconstructed networks for 175 prokaryotic genomes.

References

1. Steinmetz LM, Davis RW. Maximizing the potential of functional genomics. Nat Rev Genet 2004;5(3):190-201.

2. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2(6):418-27.

3. Young RA. Biomedical discovery with DNA arrays. Cell 2000;102(1):9-15.

4. Bulyk ML. DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol 2006;17(4):422-30.

5. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431(7004):99-104.

6. Lee TI, Rinaldi NJ, Robert F, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002;298(5594):799-804.

7. Horak CE, Luscombe NM, Qian J, et al. Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. Genes Dev 2002;16(23):3017-33.

8. Salgado H, Gama-Castro S, Peralta-Gil M, et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 2006;34(Database issue):D394-7.

9. Baumbach J, Brinkrolf K, Czaja LF, Rahmann S, Tauch A. CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. BMC Genomics 2006;7(1):24.

10. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 2004;14(3):283-91.

11. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5(2):101-13.

12. Madan Babu M, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 2006;358(2):614-33.

13. Yu H, Luscombe NM, Lu HX, et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 2004;14(6):1107-18.

14. Lozada-Chavez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 2006;34(12):3434-45.

15. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34(2):166-76.

16. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 2003;301(5629):102-5.

17. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet 2005;37(4):382-90.

18. Wang SC. Reconstructing genetic networks from time ordered gene expression data using Bayesian method with global search algorithm. J Bioinform Comput Biol 2004;2(3):441-58.

19. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 2003;19(15):1917-26.

20. Alkema WB, Lenhard B, Wasserman WW. Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. Genome Res 2004;14(7):1362-73.

21. Wang T, Stormo GD. Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A 2005;102(48):17400-5.

22. Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS. Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. Genome Biol 2004;5(11):R90.

23. Bar-Joseph Z, Gerber GK, Lee TI, et al. Computational discovery of gene modules and regulatory networks. Nat Biotechnol 2003;21(11):1337-42.

24. Xing B, van der Laan MJ. A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. J Comput Biol 2005;12(2):229-46.

25. Haverty PM, Hansen U, Weng Z. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res 2004;32(1):179-88.

26. Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics 2004;5:31.

27. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet 2001;27(2):167-71.

28. Kim H, Hu W, Kluger Y. Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. BMC Bioinformatics 2006;7:165.

29. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.

30. Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 2004;20(11):1746-58.

31. Makita Y, Nakao M, Ogasawara N, Nakai K. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. Nucleic Acids Res 2004;32(Database issue):D75-7.

32. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol 2006;360(1):213-27.

33. Balaji S, Iyer LM, Aravind L, Babu MM. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. J Mol Biol 2006;360(1):204-12.

34. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 2002;31(1):64-8.

35. Albert R. Scale-free networks in cell biology. J Cell Sci 2005;118(Pt 21):4947-57.

36. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nat Genet 2004;36(5):492-6.

37. Guelzim N, Bottani S, Bourgine P, Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 2002;31(1):60-3.

38. Lespinet O, Wolf YI, Koonin EV, Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 2002;12(7):1048-59.

39. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004;20(9):1453-4.

40. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19(2):295-6.