Journal of Biology

BioMed Central

Minireview

Reconstructing prokaryotic transcriptional regulatory networks: lessons from actinobacteria

Thiago M Venancio and L Aravind

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Correspondence: L Aravind. Email: aravind@ncbi.nlm.nih.gov

Published: 15 April 2009

Journal of Biology 2009, **8**:29 (doi:10.1186/jbiol132)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/8/3/29

Abstract

Reconstruction of transcriptional regulatory networks of uncharacterized bacteria is a main challenge for the post-genomic era. Recent studies, including one in *BMC Systems Biology*, address this problem in the relatively underexplored actinobacteria clade, which includes major pathogenic and economically relevant taxa.

Transcription regulatory networks

Since the pioneering work of Jacob and Monod [1] nearly half a century ago, which led to the operon model of prokaryotic gene regulation, genetic and molecular studies have deciphered the regulatory processes for a significant fraction of the genome of *Escherichia coli*. In the same period *Bacillus subtilis* too has risen to the status of a major model bacterium, thereby providing us with glimpses of gene regulation in two far-flung branches of the bacterial evolutionary tree. A primary outcome of these studies has been the identification of general or basal transcription factors (such as sigma factors) and specific transcription factors (such as the lac operon repressor, lacI) that together mediate the expression of target genes by binding specific regulatory DNA sequences called transcription factor binding sites (Figure 1a) [2].

Accumulation of such data in model organisms on a genomic scale has recently allowed representation of these regulatory interactions between transcription factors and their target genes as an ordered graph or a network. This transcription regulatory network provides a powerful theoretical framework to analyze the complete regulatory system of model organisms such as *E. coli* [3] or *B. subtilis* [4]. Topological studies on such networks have revealed fundamental features that are common to other biological and non-biological networks, such as an approximation of the power-law degree distribution of regulatory interactions (few transcription factors regulate many genes, and most transcription factors regulate a low number of genes) [5] and the presence of certain stereotypical recurring patterns of connections called motifs [6] (Figure 1b,c). These features are important for deciphering the responses of organisms to the environment, as well as for biochemical engineering of pathways. Three recent papers [7-9] have now reconstructed transcription regulatory networks for several species of actinobacteria.

The aftermath of the genomic revolution in biology has left us with complete genomes of numerous prokaryotes with varied ecological, economic and medical significance. However, in most of these organisms the absence of known transcription regulatory networks comparable to those assembled by classical studies in *E. coli* or *B. subtilis* is an impediment to their study and use. There has thus been

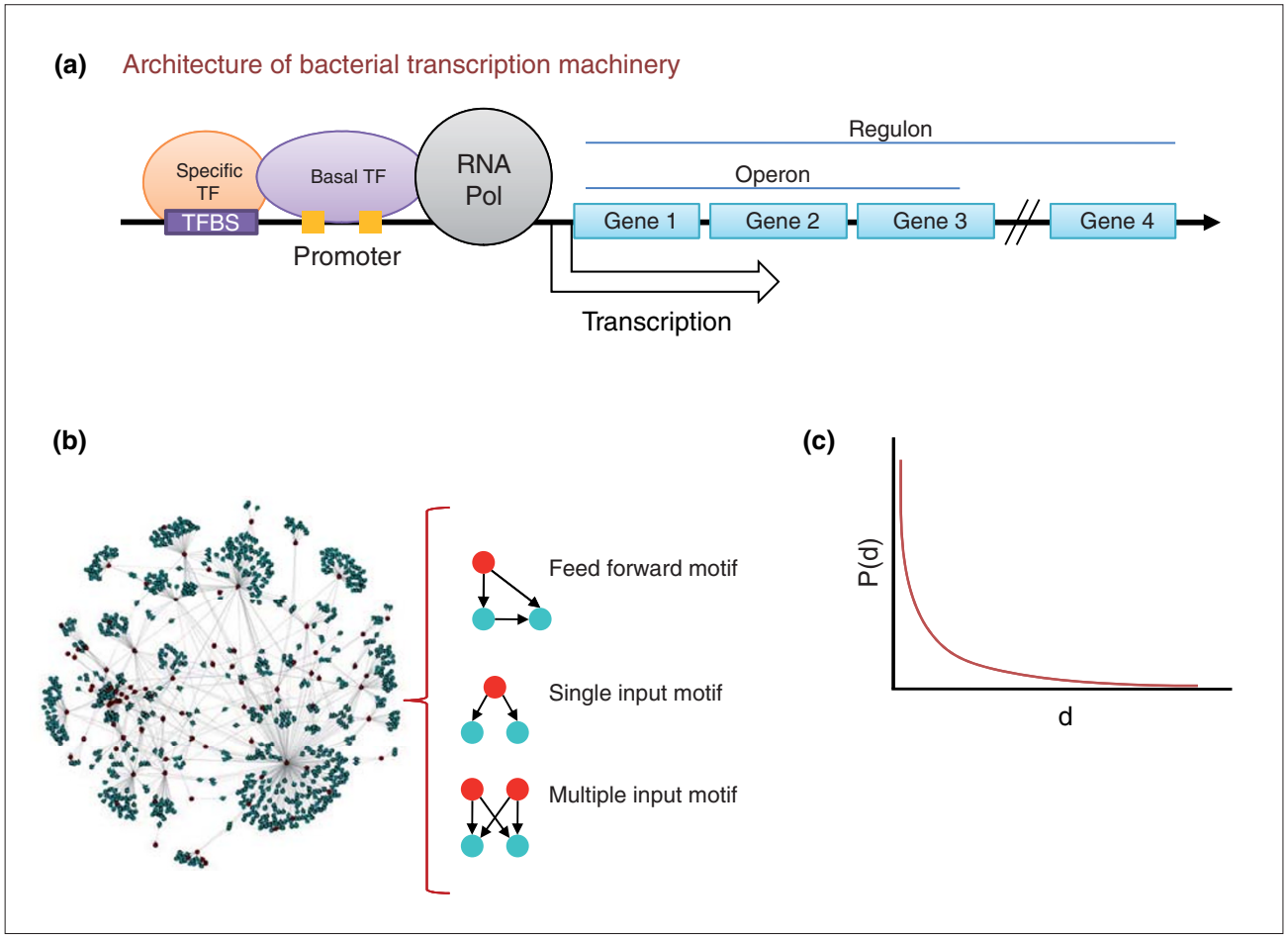

Figure 1
The transcription apparatus and transcription regulatory network of bacteria. **(a)** Schematic representation of the architecture of bacterial transcription machinery and operons and regulons. A regulon is the set of genes regulated by one transcription factor; an operon is a set of adjacent genes transcribed into one mRNA. **(b)** Architecture of transcription regulatory networks. The global structure (left) and three types of motifs found in transcription regulatory networks (right) are depicted as ordered graphs. Red dots indicate transcription factors; blue dots indicate targets. **(c)** The degree distribution of transcription factor-target interactions is approximated by a power-law equation [5]. The graph shows a power-law distribution; degree (d) is the number of regulatory connections between a transcription factor and target genes, while P(d) indicates the probability of transcription factors with a particular number of such connections. Pol, polymerase; TF, transcription factor; TFBS, transcription factor binding site.

considerable impetus to infer transcriptional regulatory interactions in organisms beyond the well studied models. Studies suggest that prokaryotic gene regulation typically takes place through certain conserved specific transcription factors operating on operons or regulons of genes, whose products are involved in well defined cellular processes (Figure 1a). Usually, these transcription factors come with a distinctive sensor domain, in addition to their DNA-binding domain, that helps them respond to the particular effector compound that induces their target regulons. These observations led to the most straightforward computational approach for reconstruction of transcription regulatory networks in uncharacterized organisms: identifying orthologs of transcription factors and target genes with respect to a template network in a model organism (such as *E. coli*) and transferring the regulatory connections to the organism of interest by assuming co-conservation of such transcription factor-target pairs (Figure 2a) [10]. An alternative approach assumes the conservation of transcription factor binding sites across distantly related prokaryotes and predicts target genes for conserved transcription factors using position-specific weight matrices or hidden Markov models derived from binding site alignments (Figure 2b).

However, both these approaches are fraught with difficulties, including the fundamental problem of correctly

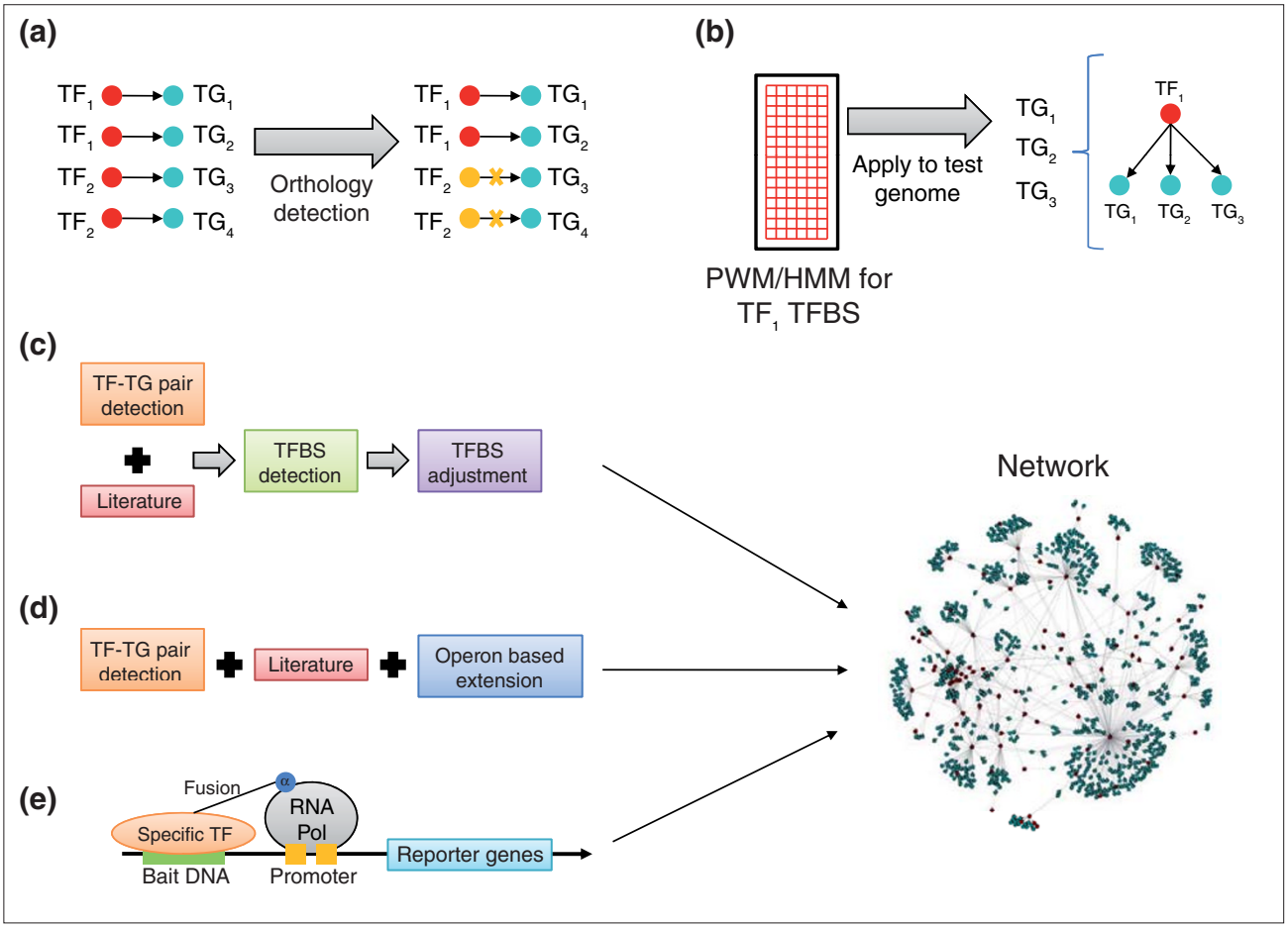

Figure 2
Methods of network inference in uncharacterized prokaryotes. **(a, b)** Conventionally used methods for network reconstruction. (a) Orthology detection by comparison of transcription factor-target (TF-TG) links between species. Crosses indicate links known from the first species that are not found in the second species. (b) Position-specific weight matrices (PWMs) or hidden Markov models (HMMs) derived from binding site alignments (represented here by a grid) are used to predict target genes for conserved transcription factors. **(c, d, e)** The three recently published approaches to network reconstruction discussed here [7-9]. (c) The approach of Baumbach *et al.* [7]. (d) The approach of Balazsi *et al.* [8]. (e) The approach of Guo *et al.* [9]. Refer to the text for details of each of the studies (c-e) aimed at reconstructing actinobacterial transcription regulatory networks. Pol, polymerase; TFBS, transcription factor binding site.

identifying orthologous transcription factors. For example, the transcription factor birA, which regulates biotin synthesis, combines an amino-terminal winged-helix-turn-helix DNA-binding domain with a carboxy-terminal biotin ligase domain. Orthologs of birA in certain bacteria lack the DNA-binding domain and thus cannot function as transcriptional regulators of biotin regulons in those organisms. Therefore, mere identification of an ortholog might not predict transcription regulation. The binding sites are usually unknown for a significant fraction of transcription factors in an organism. Even when they are known, it is observed that orthologous transcription factors can regulate orthologous targets using divergent binding sites [11], indicating the limitations of the binding-site-based approach. Furthermore, earlier studies on the relative conservation of transcription factors and targets suggest that transcription factors are more frequently displaced or lost than targets [10]. It has also been observed that the number of transcription factors encoded by a prokaryotic organism scales as a power law with respect to total gene number - larger genomes tend to have more transcriptional regulators per gene than would be expected from a linear increase with genome size. Taken together, these observations limit the scope of traditional transcription regulatory network reconstructions to well-conserved transcription factors and targets and probably work best with organisms that are phylogenetically related or are of similar genome size with a similar lifestyle [10].

New studies on network reconstruction in actinobacteria

A set of recent studies [7-9] offers new ways to tackle the challenges of reconstruction of transcription regulatory networks in uncharacterized organisms, in terms of both methodology and data. These studies focus primarily on members of the previously underexplored actinobacterial clade, including pathogens such as *Mycobacterium tuberculosis* and *Corynebacterium diptheriae* and industrially relevant organisms such as *Corynebacterium glutamicum*. The first of these, reported by Baumbach *et al*. in *BMC Systems Biology* (Figure 2c) [7], is a culmination of a series of studies on *Corynebacterium* and presents the assembly of a preliminary network for *C. glutamicum* derived from experimental results. It covers 72 transcription factors of the predicted 182 transcription factors in this organism (our unpublished results). The study [7] combined the conventional technique - detection of orthologous transcription factors and targets based on the *C. glutamicum* template - with binding site prediction to reconstruct networks in closely related uncharacterized corynebacteria: *C. diphtheriae*, *C. efficiens* and *C. jeikeium*. A key advance in this work was the adjustment of the initially inaccurately determined binding sites by shifting them by one or more positions, followed by motif searches to identify a more likely binding site. These adjusted binding sites were then used in conjunction with target gene conservation to predict actual interactions. From the results presented in this work it seems that such a dual approach, while conservative, might indeed delineate high-confidence interactions.

The second study [8] reconstructed the network of *M. tuberculosis* using a combination of experimentally documented interactions and orthology-based linkages, with an extension of these two sets of interactions using predicted operons (Figure 2d). Using this network, covering 43 of the approximately 235 transcription factors of this organism (after accounting for incorrect annotations; see below), together with microarray data, the authors were able to explore the shift in gene regulatory processes accompanying dormancy, which is a major pathogenic feature of *M. tuberculosis* [8].

The third study [9] represents a major development in terms of identification of new transcription factor-target interactions using a novel bacterial one-hybrid system. In this system, hybrid transcription factors are generated by fusing them to the α subunit of the RNA polymerase and tested for interaction with different bait DNA sequences by checking for activation of reporter genes adjacent to the bait sequence (Figure 2e). By this method the authors [9] were able to describe several novel transcription factor-target interactions related to responses to stress and redox and fatty acid metabolism in *M. tuberculosis*. Consequently, this study goes a long way in extending the network in this organism by increasing the coverage to 58 transcription factors.

A comparison of the networks from the two *M. tuberculosis* studies [8,9] showed that only ten transcription factors and nine interactions are shared. We have also assembled a transcription regulatory network for *M. tuberculosis*, using the *C. glutamicum* network reported in the Baumbach *et al*. study [7] as a template, using the conventional ortholog-based transfer of interactions (our unpublished results). This inferred network had 397 interactions, of which 49 (12.35%) were detected by either of the two studies on *M. tuberculosis* [8,9] and includes hubs that were present in both organisms, such as LexA and Crp (hubs are genes that regulate a large number of targets; LexA represses SOS-response genes and Crp is a cyclic AMP-dependent activator of gene expression). These observations strongly suggest that we are indeed far from the complete transcription regulatory network in either of these organisms. However, the independent support for about 12% of the *M. tuberculosis* interactions inferred using orthology-based techniques, even with these very incomplete networks, implies that this method has some value despite the known problems with it.

Future directions and potential pitfalls in reconstructions of transcription regulatory networks

It is sobering that these studies [7-9] still cover a relatively small fraction of the complete networks of the respective organisms. It should also be kept in mind that all of them are influenced by the state of annotation of the gene and protein databases. We noticed that in each of the studies [7-9] there are instances of false positives generated as a result of incorrect annotation of non-DNA-binding proteins as transcription factors. We further observed that most organism-specific databases do not successfully identify all potential transcription factors encoded by a particular organism. For example, most studies report the number of transcription factors in *M. tuberculosis* as ranging from 150 to 194 [8,9]. However, careful profile-based searches suggest that the actual number of transcription factors in this organism is closer to 235 (our unpublished results). Such underestimates are also observed in the case of *C. glutamicum*, suggesting that greater care needs to be applied to the detection and annotation of transcription factors.

Nevertheless, the studies [7-9] highlight some procedures that could result in improved reconstruction of transcription regulatory networks. Firstly, the success of the one-hybrid method in detecting entirely new interactions confirms that there is no substitute to an

effective high-throughput experimental method in such endeavors. This is especially true because of the presence of lineage-specific transcription factors in most bacterial clades (such as the differentiation and sporulation factor WhiB in actinobacteria), displacement of regulatory hubs (evolutionary replacement of a highly connected transcription factor in the network by another phylogenetically distinct transcription factor) and the non-linear scaling of transcription factor counts with gene number [9,10]. The *C. glutamicum* and *M. tuberculosis* network assembly efforts bring home the fact that there are already numerous individual studies in the literature that can be combined to provide a base for reconstructing a network for certain organisms. However, despite the recent progress in automatic text-mining tools [12], analysis of datasets such as those assembled in these studies [7-9] requires considerable human intervention to generate reliable transcription-factor-target connections. Finally, the novel approach of combining transcription factor-target orthology with adjusted transcription factor binding site predictions presented in the corynebacteria study [9] serves as a plausible model for making reliable predictions of interactions, at least for closely related taxa. This, in conjunction with high-throughput experimental studies targeting representatives across the prokaryotic tree, might indeed prove useful in future efforts towards accurate transcription regulatory network reconstruction.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. We thank Jan Baumbach for the assistance in obtaining the *C. glutamicum* transcription network.

References

1. Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3:**318-356.
2. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, 61:1053-1095.
3. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J: **RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation.** *Nucleic Acids Res* 2008, **36:**D120-D124.
4. Sierro N, Makita Y, de Hoon M, Nakai K: **DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information.** *Nucleic Acids Res* 2008, **36:**D93-D96.
5. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, 5:101-113.
6. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of *Escherichia coli*.** *Nat Genet* 2002, **31:**64-68.
7. Baumbach J, Rahmann S, Tauch A: **Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms.** *BMC Syst Biol* 2009, **3:**8.
8. Balazsi G, Heath AP, Shi L, Gennaro ML: **The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest.** *Mol Syst Biol* 2008, **4:**225.
9. Guo M, Feng H, Zhang J, Wang W, Wang Y, Li Y, Gao C, Chen H, Feng Y, He ZG: **Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system**. *Genome Res* 2009, doi:10.1101/gr.086595.108.
10. Madan Babu M, Teichmann SA, Aravind L: **Evolutionary dynamics of prokaryotic transcriptional regulatory networks.** *J Mol Biol* 2006, **358:**614-633.
11. Mazon G, Lucena JM, Campoy S, Fernandez de Henestrosa AR, Candau P, Barbe J: **LexA-binding sequences in Gram-positive and cyanobacteria are closely related.** *Mol Genet Genomics* 2004, **271:**40-49.
12. Rodriguez-Penagos C, Salgado H, Martinez-Flores I, Collado-Vides J: **Automatic reconstruction of a bacterial regulatory network using natural language processing.** *BMC Bioinformatics* 2007, **8:**293.