Cell. Mol. Life Sci. 63 (2006) 2260–2290
1420-682X/06/202260-31
DOI 10.1007/s00018-006-6184-6

Cellular and Molecular Life Sciences

Review

Global analysis of gene transcription regulation in prokaryotes

D. Zhou* and R. Yang*

State Key Laboratory of Pathogen and Biosecurity, National Center for Biomedical Analysis, Institute of Microbiology and Epidemiology, Academy of Military Medical Sciences, 20, Dongdajie, Fengtai, Beijing 100071 (China), e-mail: dongshengzhou1977@gmail.com; ruifuyang@gmail.com

Received 24 April 2006; received after revision 30 May 2006; accepted 15 June 2006
Online First 21 August 2006

Abstract. Prokaryotes have complex mechanisms to regulate their gene transcription, through the action of transcription factors (TFs). This review deals with current strategies, approaches and challenges in the understanding of i) how to map the repertoires of TF and operon on a genome, ii) how to identify the specific cis-acting DNA elements and their DNA-binding TFs that are required for expression of a given gene, iii) how to define the regulon members of a given TF, iv) how a given TF interacts with its target promoters, v) how these TF-promoter DNA interactions constitute regulatory networks, and vi) how transcriptional regulatory networks can be reconstructed by the reverse-engineering methods. Our goal is to depict the power of newly developed genomic techniques and computational tools, alone or in combination, to dissect the genetic circuitry of transcription regulation, and how this has the tremendous potential to model the regulatory networks in the prokaryotic cells.

Keywords. Prokaryote, gene transcription, transcription factor, operon, regulon, regulatory network, microarray expression profiling, ChIP-chip.

Transcription factors in prokaryotic gene regulation

Regulation of gene transcription at promoters

In transcriptional regulation in prokaryotes, expression of a gene is controlled at the stage of RNA synthesis by a regulator that interacts with a specific regulatory DNA element. Synthesis of RNA is under the direction of DNA by the RNA polymerase enzyme (Fig. 1). RNA polymerase consists of the core enzyme and the sigma factor. A RNA core polymerase is a multi-subunit complex with a general structure of $\alpha_2\beta\beta'$ that undertakes the elongation of RNA [1]. Sigma factor is needed for the initiation of RNA transcription, and it is a major influence on selection of promoters [2].

Transcription factor (TF) is a protein needed to activate or repress the transcription of a gene, but is not itself a part of the enzymes [3–5]. Some TFs bind to cis-acting DNA sequences only; some bind to each other; others bind to DNA as well as to other TFs [3–5]. Regulation of gene transcription in an organism involves a complex network, where the DNA-binding TFs are a key component. They regulate the transcription of specific genes by acting on the cis-regulatory sequence (TF-binding sites) within the promoters of these genes (Fig. 1). Based on sequence and structural homologies, DNA-binding regions of the prokaryotic TFs have been assigned to a number of families of DNA domains [6, 7], including the three most well characterized ones, the helix-turn-helix, the winged helix and the β ribbon [8].

Transcription activators and repressors

When a TF binds to a specific promoter, it can either activate or repress transcription initiation [4, 5]. An activator stimulates the expression of its target gene, typically

* Corresponding authors.

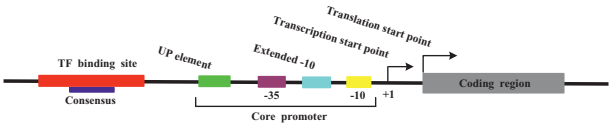

Figure 1. Structure of a prokaryotic promoter. A promoter is a region of DNA on the genome where RNA polymerase and TF bind to initiate transcription. The +1 indicates the base pair where transcription initiates, and it is commonly called transcription start point. Base pairs upstream of the transcription start point are assigned positive numbers, while those downstream are shown with negative numbers. The core promoter consists of –10, –35 and extended –10 and UP elements. The –10, –35 and extended –10 elements are recognized by domains 2, 4 and 3 of the RNA polymerase σ subunit, respectively [1, 246]. The UP element, located upstream of the –35 element, is recognized by the C-terminal domains of the RNA polymerase α subunits [247]. Sometimes, the TF binding site may overlap the core promoter sequence. A consensus sequence is often located in the TF-binding site.

by acting on a promoter to stimulate RNA polymerase. For negative control, the TF is a transcription repressor that either binds to DNA to prevent RNA polymerase from initiating transcription, or binds to messenger RNA (mRNA) to prevent a ribosome from initiating translation. Some TFs function solely as activators or repressors, whereas others can function as either (dual regulators) according to the target promoters. A computational analysis of *Escherichia coli* K-12 genome estimates a total of 314 TFs that consist of 35% activators, 43% repressors and 22% dual regulators [9].

Global transcription regulators

Global transcription regulators are TFs (i) that have the ability to regulate large numbers of genes that belong to different functional classes, (ii) that control a complex regulatory cascade by a mechanism of not only directly controlling the expression of specific genes, but also indirectly regulating various cellular pathways by acting on a set of local regulators controlling just one or a few genes, and (iii) that act on the target promoters that use different sigma factors [10]. This definition excludes TFs involved in essential cellular functions [10]. It has been estimated that seven global transcription regulators (CRP, FNR, IHF, Fis, ArcA, NarL and Lrp) in *E. coli* control 50% of all regulated genes, whereas ~60 TFs each control only a single promoter [10].

Virulence-related transcription factors

During infection a pathogen is exposed to a series of environmental changes that can make its living conditions far from optimal. To survive the stressful environments, pathogens must make appropriate adaptive and/or protective responses, primarily reflected by transcriptional changes in specific sets of genes. Expression of virulence determinants, which allows pathogens to multiply on and within host cells and tissues, are tightly and coordinately regulated during specific stages of infection [11]. Regulation of virulence genes is no exception in involving TF-DNA interactions. Virulence-related TFs can sense host signals such as changes in temperature, osmolarity, pH, iron levels, nutrient availability, antimicrobial agents and oxygen levels, etc. [12–18]. In addition to stimulating the expression of virulence genes that can actively attack host defense mechanisms, these TFs still differentially regulate other broad sets of genes, which is required for adaptation to host niche [12–18]. Disruption of these TFs results in reduced virulence of the mutants due to disordered transcriptional responses of the pathogens during infection.

Identification and characterization of transcription factors

Genome-wide prediction of transcription factors

Identification of DNA-binding TFs is crucial to understanding gene regulatory mechanisms. Preliminary TF-encoding information on a sequenced genome comes from genome annotation by detecting factors homologous to known TFs [19], or by functional classification schemes that assign proteins to the category of transcription regulation [20]. More sophisticated TF prediction methods are based on computational collection and assignment of DNA-binding motifs, enabling genome-wide TF prediction for the model microorganism *E. coli* [9, 21] and even for organisms from across the tree of life [22, 23].

Based on determination of the homology between the domains and protein families of the TFs and their regulated genes, and proteins of known three-dimensional structure, a computational method has been established to identify what is likely to be the large majority of *E. coli* TFs [21]. In this approach, 11 families of DNA-binding domains are identified from public databases. Subsequent assignment of these superfamilies to *E. coli* proteins generates a preliminary set of 416 proteins with DNA-binding domains. After removing proteins involved in transposases and replication/repair and other enzymes, a final set of 271 TFs is obtained.

Doerks et al. [22] present a method that exemplifies how genomic context searches work to identify TFs from a wide variety of prokaryotic species with available whole-genome sequences. The authors first extract clusters of orthologous groups (COGs) involved in transcription regulation from the COGs database [24]. Enzyme-related COGs are subsequently removed. Each of the resulting 128 groups contains orthologous TFs derived from several genomes. When these COGs of known and putative TFs are projected to *E. coli* K-12, they cover 85% of the list of *E. coli* TFs described in [21].

A procedure [23] that uses profile hidden Markov models (HMMs) of domains from the SUPERFAMILY [25] and Pfam [26] databases is proposed to automatically predict DNA-binding TFs. Using powerful multi-sequence comparison, HMMs recognize only TFs that use the mechanism of sequence-specific DNA binding. This method is applied to more than 150 completely sequenced genomes from across the three domains of life, leading to the establishment of a comprehensive TF database, DBD [24].

DNA pull-down strategies

There is a big gap between the promoter DNA elements and the predicted TFs scattered over the prokaryotic genomes. In many circumstances, binding factors for a promoter of interest are unknown. DNA pull-down strategies, including DNA affinity chromatography and gel mobility shift assay, are successful in isolation and identification of sequence-specific DNA-binding factors from nuclear extracts.

Because of its high selectivity, DNA affinity chromatography (Fig. 2), is the most widely used technique for purification of TFs and other DNA-binding proteins [27, 28]. The isolated DNA-binding proteins are subsequently separated on SDS-polyacrylamide gel electrophoresis (PAGE), and their identities are determined by mass spectrometry (MS) [29, 30]. Various affinity supports, such as agarose, Sepharose, cellulose and silica, are routinely used for coupling DNA, and a wealth of coupling chemistries are available for attaching DNA to these supports [28, 31]. Conventional DNA affinity chromatography is quite laborious and time-consuming. Further modifications and improvements have been widely proposed [32–37].

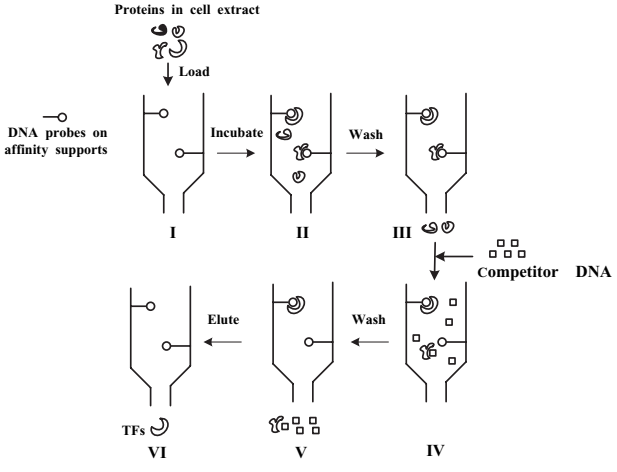

Figure 2. DNA affinity purification of TFs. DNA probes containing a TF-binding site are either adsorbed or linked covalently to a chromatographic support. The nuclear or whole cell extract as a rich source of TFs is incubated with DNA probes, and the corresponding TFs specifically bind to the DNA (steps I and II). The subsequent washing can remove most other proteins, rather than the DNA-binding TFs and some contaminant proteins that bind to the DNA probes weakly and nonspecifically (step III). When a sufficient amount of competitor DNA such as poly(dI-dC) is added, the weaker binding proteins will bind this competitor (step IV) and are then washed out (step V). The TFs specifically binding to the DNA probes are finally eluted under the stringent conditions (step VI).

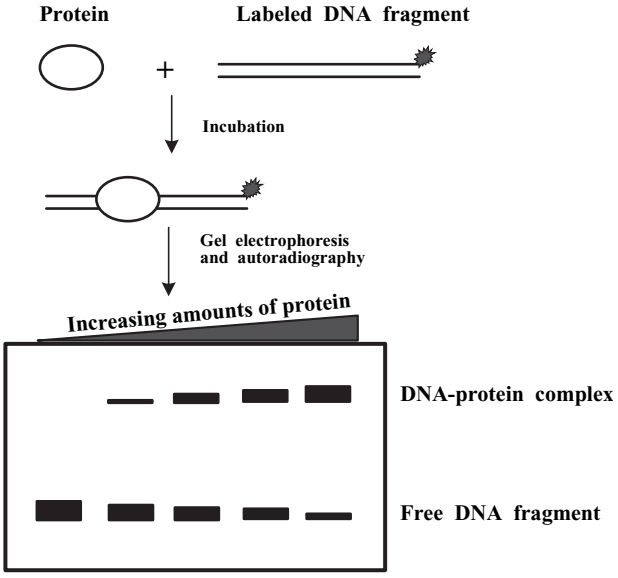

Figure 3. Electrophoretic mobility shift assay. Increasing amounts of TF sample are incubated with a radiolabeled DNA fragment. The reaction products are then analyzed with the non-denaturing PAGE. The distribution of radioactivity is viewed by radioautography. DNA molecules to which TFs bind move more slowly in the gel and are retarded relative to the sample with no protein.

In the electrophoretic mobility shift assay (EMSA) (Fig. 3), a radiolabeled specific DNA is incubated with cell extract, and the mixture is then subjected to non-denaturing PAGE. If the corresponding DNA-binding TF is present in the cell extract, it retards the mobility of the probe on PAGE, which can easily be detected by autoradiography. When a specific antibody against a candidate TF is available, a supershift is observed because of formation of DNA-TF-antibody complex.

Conventional EMSA is restricted to candidate TFs and the availability of the specific antibodies. However, if no candidates can be proposed, EMSA is of limited utility in identification of novel DNA-binding TFs. Woo et al. [38] present a method for the identification of DNA-binding proteins seen in EMSA using the power of two-dimensional electrophoresis coupled with mass spectrometry. The method consists of four phases. First, nuclear proteins are partially purified by S300 gel filtration. The MM and pI of the protein are then estimated by coupling SDS-PAGE or IEF (isoelectric focusing) with EMSA. Next, gel slices are excised from a two-dimensional gel at the predetermined pI and MM coordinate. Proteins are eluted, re-natured, and tested for DNA-binding activity in EMSA. Identified protein spot candidates are subjected to MS to determine their identity. Hazbun et al.

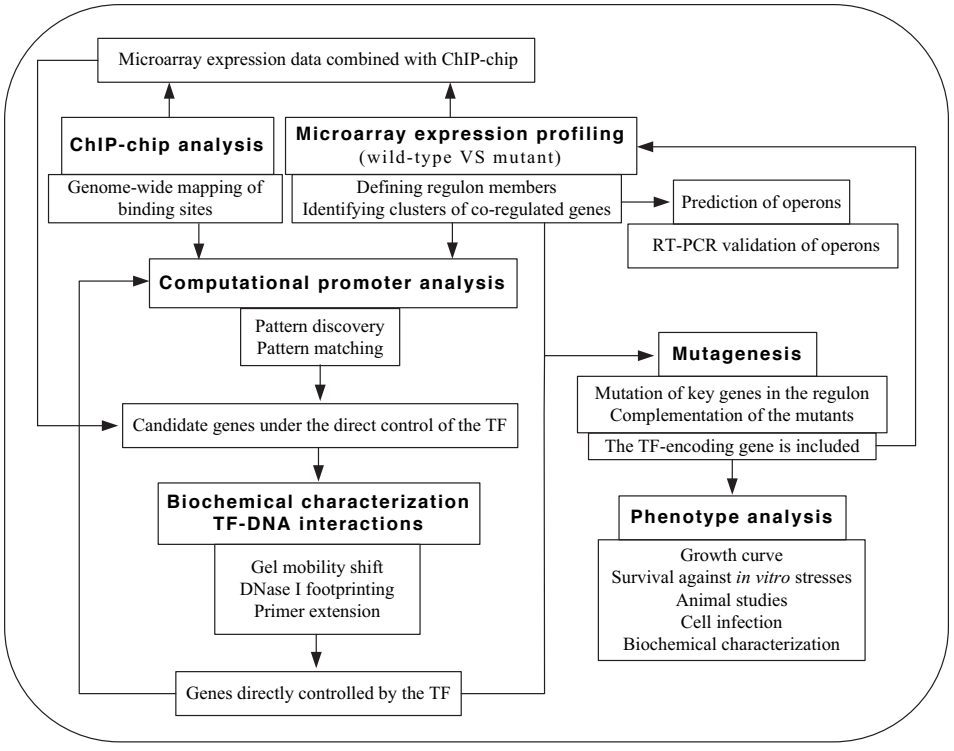

Figure 4. Characterization of a TF of interest with TF pull-down strategies. This figure indicates the TF pull-down strategies aiming to give a comprehensive functional characterization of a specific TF.

[39] report the use of a genome-wide EMSA to identify proteins capable of binding to a cis-acting regulatory element. Using *Saccharomyces cerevisiae* as model system, they prepare an array of 6144 yeast strains, each overexpressing the single yeast open reading frame (ORF) fused to glutathione S-transferase (GST). Protein pools are then generated by purification of the GST fusion proteins from whole cell extracts from different groups of strains. Each protein pool is used in an EMSA to detect the binding of proteins to a radiolabeled DNA fragment. This report demonstrates the feasibility of genome-wide screening of proteins for binding to a specific regulatory DNA of interest by rapidly assaying a large fraction of ORFs of an organism.

Transcription factor pull-down strategies

So far, a dozen families of TFs have been identified in prokaryotes [9, 21], including the well-characterized AraC [40], CRP [41], LacI [42], Lrp [43], LysR [44] and MerR [45] families. Based on their ability to recognize and interact with specific regulatory DNA sequences present in the promoters, TFs along with their target genes constitute complex regulatory networks involved in both normal cell growth and survival against stress or host defense. Thus, understanding the role of TFs in maintaining and altering expression levels of their target genes, as well as the phenotypic characteristics therein, is crucial to understanding normal cellular function as well as disease. Figure 4 shows the TF pull-down strategies for characterization of a TF of interest, which will be discussed one by one below.

Microarray expression profiling

Two-sample co-hybridization experiment

DNA microarray is able to determine changes in mRNA levels simultaneously for all the genes in a cell. In a typical two-sample experiment (Fig. 5), RNA is extracted from reference and test samples, respectively, labeled with different fluorescein dyes, and co-hybridized to a complementary DNA (cDNA) microarray. The hybridized microarray slides are scanned, and data extracted from microarray images are subjective to exclude poor-quality spots [46, 47]. In general, spots with background-corrected signal intensity in both channels less than two-fold background intensity are removed from further analysis. The resulting data set is subsequently normalized through balancing the fluorescence intensities of the two labeling dyes. Normalization serves to remove the systematic variations in the measured gene expression levels of two co-hybridized samples, so that biological differences can be more easily distinguished [48–51]. The systematic variations in microarray experiments come from differences in the number of cells in the cultures, RNA

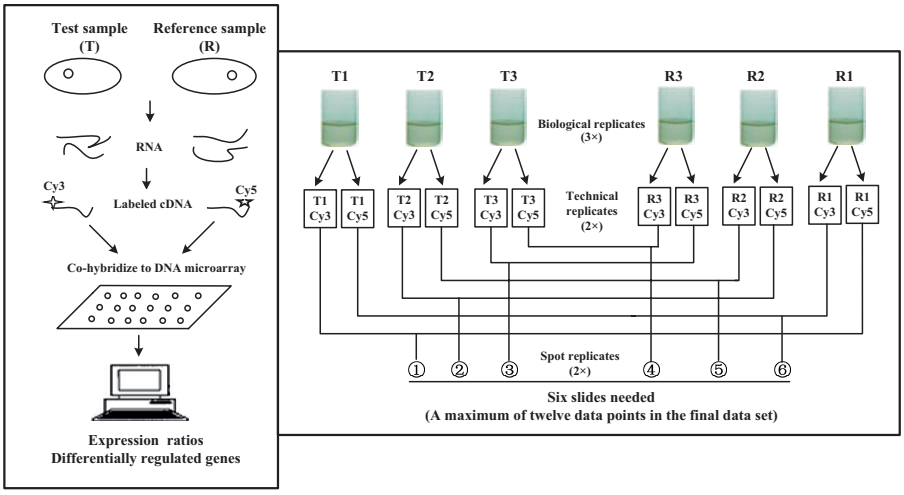

Figure 5. Designs for the two-sample experiment. The left part of this Figure shows a typical two-sample experiment, where total cellular RNA is extracted from reference and test samples, respectively. RNA samples are reverse-transcribed into cDNA with attendant incorporation of different fluorescein dyes, usually a red-fluorescent cyanine 5 (Cy5) and a green-fluorescent cyanine 3 (Cy3). A mixture of differently labeled cDNA samples hybridizes to a whole-genome cDNA microarray. The right side depicts the three types of replication for a single microarray experiment: biological replicates (independent cell cultures), technical replicates (separated microarray slides) and spot replicates (genes spotted in duplicate on each slide).

extraction efficiency, dye-labeling efficiency, hybridization efficiency, heat and light stability of dyes, scanning properties, and scanner settings for the two channels.

The commonly used normalization methods include total mRNA normalization, which uses all genes on the microarray [50, 52], housekeeping normalization using genes with invariant expression [53], external spike-in control normalization, which uses a known amount of exogenous control genes added during hybridization [54, 55] and the nonlinear locally weighted scatterplot smoothing (LOWESS) normalization [49, 56]. To compare the mRNA profiles between reference and test samples, the averaged expression ratio of test/reference for each gene is calculated and then logarithm-transformed usually to base 2. Using the logarithm has the advantage of producing a continuous spectrum of values and treating up- and downregulated genes in a similar fashion [50].

A fixed threshold cutoff method (e.g. a twofold increase or decrease in gene expression) is not sufficient to identify differentially regulated genes, given the reasons that a gene with low expression in one or both strains has more variable expression ratios than a gene with a more substantial level of basal expression [57], and that nonsystematic variations (e.g., random biological variations, sample handling errors and measuring errors) cannot be handled by data normalization [58, 59]. Random biological variations come from the physiological differences in growth microenvironments in cultures (e.g., nutrients and temperature), growth phase and multiple additional stochastic effects that cannot be controlled. It has been reported that even when bacterial cells grown under two 'identical' conditions are compared with each other, differences in gene expression are still observed [60].

Significant changes of gene expression are commonly identified on the basis of replicate microarray data. Replication of a microarray experiment is essential, as it gives a baseline to measure the non-systematic variations in statistic calculation [61]. There are three types of replication (Fig. 5). First, total RNA is extracted from independent cell cultures (biological replicates). Second, various aliquots of each RNA extraction are used to prepare the labeled probes for separated microarray slides, for which (technical replicates) the incorporated dye is reversed (dye swaps). Third, each gene or ORF is present in duplicate on the printed slides (spot replicates). Technical replicates for two separated microarray slides (Fig. 4) come from the same RNA extraction (the same biological replicate). Dye swaps are designed for these two technical replicates. On one slide the test sample is assigned to Cy5 and the reference sample is assigned to Cy3, while on another slide the dye assignments are reversed. Data normalization is not likely done equally well for every spot on every slide, so there may be a residual dye bias. Averaging dye-swap data will make an experiment less prone to this kind of dye bias [62].

The commonly used statistical methods for discovering differentially expressed genes include standard or regularized two-sample *t*-test [63–65], ANOVA (analysis of variance) and its variants [66–68], and the maximum likelihood [61, 69] and mixture models [70, 71] (Table 1). The shared features of these methods are that they rank the genes in order of evidence, from strongest to weakest, for differential expression, and that they can assess the rate of false positives (unchanged genes declared differentially expressed) and rate of false negatives (missed differentially expressed genes) [72–74].

Table 1. Selected leading tools for microarray expression data analysis.

Tool	Description	Reference	URL
Image processing			
ScanAlyze	Semi-automatic definition of grids and complex pixel and spot analyses.		http://rana.lbl.gov/EisenSoftware.htm
GenePix Pro	Commercial softwares provided with the microarray scanners for spot identification, data extraction, scatter plot, histogram and data normalization.		http://www.moleculardevices.com
QuantArray			http://www.packardbioscience.com/
Identifying differentially expressed genes			
SAM	SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, the false discovery rate (FDR).	[64]	http://www-stat.stanford.edu/%7Etibs/SAM/index.html
Cyber T	Cyber-T employs statistical analyses based on simple *t*-tests that use the observed variance of replicate gene measurements across replicate experiments, or regularized *t*-tests that use a Bayesian estimate of the variance among gene measurements within an experiment.	[65]	http://visitor.ics.uci.edu/genex/cybert/
EDGE	EDGE can be used to perform significance analyses for both two-sample and time-course experiments. This approach is based on the 'optimal discovery procedure' (ODP) that uses all relevant information from all genes in order to test each one for differential expression.	[109]	http://faculty.washington.edu/jstorey/edge/
Onto-Express	Onto-Express is able to automatically translate differentially repressed genes into functional profiles, using Gene Ontology.	[253]	http://vortex.cs.wayne.edu/Projects.html#Onto-Express
Clustering			
Cluster and TreeView	Cluster, one of the most widely used clustering tools, performs a variety of types of cluster analysis, including hierarchical clustering, SOMs, k-means clustering and PCA. TreeView graphically browses results from Cluster.	[82]	http://rana.lbl.gov/EisenSoftware.htm
STEM	STEM is specific for clustering, comparing and visualizing short time series gene expression data.	[90]	www.cs.cmu.edu/~jernst/stem/
Software package			
TM4	A package of open source software programs composed of MicroArray DAta Manager (MADAM), Spotfinder, Microarray Data Analysis System (MIDAS) and MultiExperiment Viewer (MEV).	[254]	http://www.tigr.org/software/microarray.shtml

The subsequent step is to choose a cutoff value for the ranking statistic to pick out genes considered as significant, given that only a limited number of genes will be differentially expressed in a typical two-sample experiment. For instance, the standard *t*-test produces a *p*-value that represents the probability of difference observed. A very small *p*-value indicates that the tested gene is likely to be differentially expressed. Depending on the percentage (e.g., 5%) of false positives chosen, an appropriate threshold (e.g., $p < 0.05$) can be selected to pick out the genes differentially expressed.

Identifying stimulons and regulons

A stimulon is a group of genes or operons that are differentially expressed in response to a given environmental perturbation [75, 76]. To define the stimulon, the cDNA microarray is used to compare gene expression patterns in wild-type (WT) strain under a stimulating condition (test sample) with an unperturbed control (reference sample). DNA microarray-based stimulon studies will clarify the vigilance of an organism to environmental changes and the alacrity of the transcriptional response, giving a global perspective allowing one to see that seemingly

unrelated activities are modulated together. It provides numerous new avenues for focused hypothesis-based investigations to delineate the role(s) of specific genes or operons in environmental response and adaptation, thus indicating the nature and function of signaling pathways activated upon specific environmental changes.

A regulon includes all target genes controlled either directly or indirectly by a single TF [75, 76]. For identifying regulon members, RNA from cells expected to have low or no expression of regulon genes is compared with RNA from cells substantially expressing regulon members. The standard procedure is the comparison of expression profiles between a WT strain (reference sample) and the isogenic mutant (test sample) of a TF. Genes with differential expression are considered as regulon members controlled by the TF involved in mutation. The function of a TF often relies on its ability to sense specific environmental conditions. Therefore, microarray experiments are carried out by use of media conditions known to be important for the TF to trigger transcriptional pathways. Growth conditions for a large set of bacterial TFs are stored in the RegulonDB database [77]. Most of these conditions have historically been used for *in vitro* stress studies and are thus suboptimal for normal bacterial growth; these stimulating conditions (environmental perturbations) are often considered as the host-responding signals during pathogenic infection.

Clustering analysis of multiple two-sample experiments

A collection of multiple two-sample experiments will generate a matrix of expression ratios, with genes in rows and conditions in columns. Thus, each column represents a single two-sample experiment. The expression level of a gene over conditions is called a gene expression profile. Subsequent clustering analysis will identify clusters of genes with similar expression profiles. Expression profiles within a cluster are more similar to each other than those in different clusters.

Clustering can be viewed as a data reduction process, in that observations of gene expression in each cluster can be over-represented. This process will produce much greater insight into functional classes of co-expressed genes, since genes functionally related, i.e. belonging to the same regulatory pathway or to the same functional complex, should be co-regulated and consequently should show similar expression profiles. Thus, the clustering genes with similar expression profiles can potentially be utilized to predict the functions of gene products with unknown functions, and to identify sets of genes that are co-expressed to play the same roles in the cell cycles.

Various clustering algorithms either supervised or unsupervised have been successfully applied to microarray expression data [78–81] (Table 1). The unsupervised methods include hierarchical clustering [82], *K*-means clustering [83], self-organizing maps (SOMs) [84] and principle component analysis (PCA) [85], all of which calculate pairwise distances or similarities between pairs of gene expression profiles in the process of clustering. Unsupervised methods attempt to detect natural groups of co-regulated genes in microarray data, unbiased by outside knowledge. They require no additional knowledge or classification scheme besides the expression data themselves. An alternative for identifying patterns of gene expression is the supervised methods, if one has some previous information about which genes are expected to cluster together [86]. Supervised methods require pre-existing classification information deriving from outside microarray experiments. One of the widely used unsupervised methods is the support vector machine [87].

Time-course experiment

In the two-sample experiment, differences in gene expression are measured at a single time point. Thus, differential expression is studied from a static viewpoint. The regulation of gene expression is a dynamic process, so it is also important to characterize changes in gene expression over time. Typically, gene expression levels are compared across a number of time points. An important issue in the time-course experiment is the design of sampling rates. If the experiment is undersampled, the results might not correctly represent the activity of the TFs in the duration of the experiments, and key events will be missed; on the other hand, oversampling is expansive and time consuming [88]. It must be borne in mind that action of the TF under *in vitro* stimulating conditions for an overly long time might result in regulatory concentrations exceeding normal titers. In this situation, the TF can occupy sequence-proximate but physiologically irrelevant sites, or related sites normally bound by another TF, which will bring the incorrect assignment of irrelevant genes as regulon members [75].

A time-course experiment also generates a matrix of expression ratios, with genes in rows and time points in columns. The clustering algorithms described above treat their input as a vector of independent samples, i.e., they assume that data at each time point are collected independent of each other. They ignore the time sequence and the time dependence of the data between time points. In addition, most of the gene expression time series come from an unknown distribution. Therefore, conventional clustering methods appear to be less appropriate for such data. Although there are gene expression time-course experiments with as many as 80 time points [89], the majority of time series are much shorter. A survey of the Stanford Microarray Database (SMD) shows that more than 80% of the available time-course datasets contain ≤ 8 time points [90], and thus the resulting data are prone

to contain different kinds of non-idealities. More recently, a number of clustering algorithms were specifically designed for microarray expression data of short time course [90–96] (Table 1) as well as relatively long time course [97–100].

In the simple two-sample approach, expression ratios are collected from various replicates that belong to a single 'group' and without respect to time course. The task of the time-course experiment is to find changes in gene expression at different times. Clustering analysis contribute nothing to this process. Recently, algorithms has been proposed to exploit information in the time-course gene expression data to detect statistically significantly periodically expressed genes [101–109] (Table 1). Because of the timing of the genetic response, primary target genes for the TF may be those whose expression changes first, whereas those that are indirectly affected will be modified later [75, 110]. Thus, differentially expressed genes in time-course experiments represent both first- and second-order downstream effects of the disrupted TF, which in turn can be used to identify target genes and to construct regulatory networks.

Validation of microarray data

Microarray results are influenced by microarray construction, RNA extraction, probe labeling, hybridization conditions and data analysis [111, 112]. Because of the inherent limitations in reliability, microarray results should be validated with at least one traditional methods such as Northern blot, polymerase chain reaction (PCR), and *lacZ* reporter fusion [112, 113].

Northern blot represents the oldest method for detection of specific mRNA based on hybridization to labeled gene-specific probes. One of the big defects of this technology is that it is not very sensitive. The difficulty of getting large enough amounts of RNA has discouraged wide utilization of this technology at present. PCR appears to be the method of choice as it is rapid and requires a minimal starting template. The same source of RNA used in the primary microarray expression analysis should be used in reverse transcriptase (RT)-PCR validation experiments [114]. For conventional RT-PCR, there is no reliable linear relationship between the amount of starting template and the amount of product formed after a fixed number (e.g. 30) of cycles, unless the reaction is proceeding exponentially at the time point of detection. Real-time RT-PCR using fluorescent reporter molecules has its own way of monitoring production of amplification products during each cycle of the PCR reaction [115, 116]. Either gene-specific anti-sense primers or random hexamers can be used to probe cDNA synthesis. For bacteria, mRNA transcript is not polyadenylated at its 3′ terminus. There may be rapid mRNA decay initiated by endonucleolytic cleavage followed by 3′-to-5′ exonucleolytic degradation. In this situation, random hexamers are preferred for extension of cDNA. To compare mRNA levels between reference and test samples, expression ratios should come from the same starting amount of total mRNA. Therefore, normalization is conducted by carrying out a parallel determination of another gene (a 'housekeeping' gene) that is transcribed at the same level in the two samples [117]. In many cases, this gene is unknown. One has to use genes whose transcription is identical between the two samples as determined by both microarray and real-time RT-PCR [118, 119]. When a large number of genes are subject to RT-PCR, construction of an absolute standard curve for each gene with serial dilutions of known template is laborious. Alternatively, the relative standard curve is simply constructed with a single gene with a high mRNA level identified by microarray analysis, using serial dilution of cDNA prepared from one sample [114]. Some investigators choose dozens of genes exhibiting high, moderate and low change in expression (as determined by microarray) to compare data from real-time RT-PCR and microarray [118, 120]. The resulting logarithm-transformed expression ratios from real-time PCR are plotted against those obtained by microarray analysis. A strong positive correlation between the two techniques indicates the reliability of microarray data.

Reporter genes are widely used as 'markers' for analysis of up- and downregulation of gene expression [121]. One of the most common reporter genes used is the *E. coli lacZ* gene, which codes for an active subunit of β-galactosidase [122]. One can start by cloning of a fragment of DNA upstream of a gene or an operon identified by microarray, using a plasmid vector carrying the promoterless *lacZ* reporter gene (Fig. 6). The recombinant vector containing the promoter sequence is subsequently transformed into mutant and WT, respectively. The β-galactosidase expression can be easily measured by its catalytic hydrolysis activity of O-nitrophenyl-β-D-galactopyranoside substrate to a bright yellow product. The β-galactosidase activity should be proportional to the rate of transcription of the gene or the operon whose upstream regulatory DNA fragment is cloned upstream of *lacZ*. This assay will ultimately demonstrate whether the promoter activity of a DNA fragment is under the control of the TF involved in mutation (see examples in [123]). An alternative is detection of β-galactosidase with the fluorescein di-β-D-galactopyranoside substrate, which has been shown to be several orders of magnitude more sensitive [124]. It should be noted that simple fusion of promoter DNA into the reporter plasmid has inherent problems, such as disordered promoter activity of the cloned DNA fragment, titration of TFs due to the copy number of the plasmid, read-through of endogenous plasmid promoters and growth phase-dependent alteration of plasmid copy number [125]. Rather than introducing it into the recombinant plasmid, single-copy *lacZ* fusion can be introduced

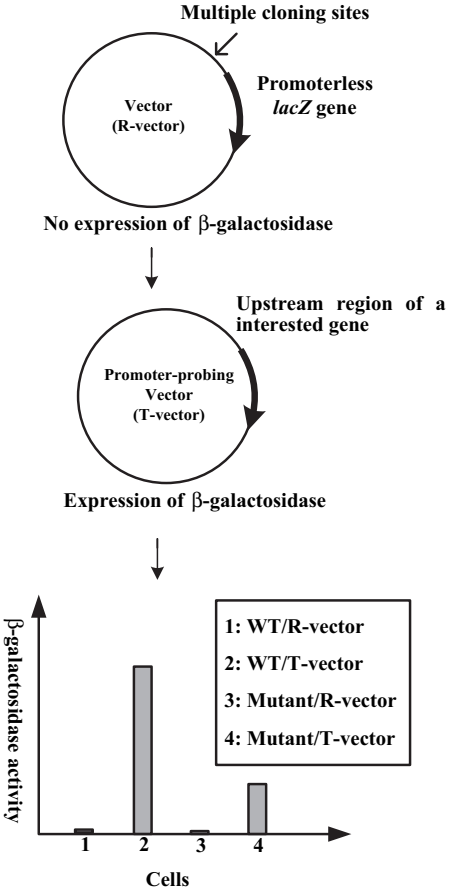

Figure 6. *LacZ* reporter fusion. A promoter DNA fragment presumably dependent on a TF is cloned into a plasmid vector carrying the promoterless *lacZ* reporter gene (R-vector). The recombinant vector (T-vector) is subsequently transformed into a mutant of the TF and its isogenic WT strain, respectively. The detecting *β*-galactosidase activity indicates the promoter activity of the cloned DNA fragment under the control of the TF.

into a specific chromosomal position by site-specific homologous recombination [126–129], for which the *lacZ* reporter is usually inserted downstream of a gene of interest such that the WT coding sequence is maintained.

Prediction and identification of operons

Structure of prokaryotic operons

In prokaryotes, an operon consists of one or more genes which are transcribed to a single polycistronic RNA transcript, as well as the regulatory elements recognized by regulator(s) (Fig. 7). An upstream promoter and a downstream terminator delimit an operon, and usually no promoter or terminator can be found within the operon. Genes in an operon, commonly functionally related, are separated by a short length of DNA and arranged in tandem in the same orientation on the same strand of a genomic sequence. Organization of operons on prokaryotic

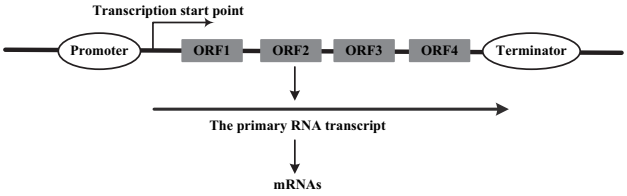

Figure 7. The structure of an operon. RNA is transcribed from the translation start site located between the promoter and the start codon. RNA polymerase moves along the template, synthesizing RNA, until it reaches a terminator sequence. It may include more than one gene. The primary transcript is the original unmodified RNA product consisting of a leader, a tail, coding regions and spacers (if polycistronic). The polycistronic primary transcripts are clipped to remove the leader, trail, and spacers, and to give the separate, mature mRNA products.

genomes is believed to facilitate the efficient coordinated regulation and association of functionally related protein products. Operons represent a basic organizational unit in a highly compartmentalized and hierarchical structure of cellular processes in a cell.

Whole-genome prediction

Characterization of operons certainly provides the basic knowledge to reconstruct biological pathways and the regulatory networks. A number of computational methods have been developed for operon prediction from genomic sequences. In the majority of these methods, statistical models are generated through training with experimental information (distance of adjacent genes, transcription orientation gene order, promoters and terminators, etc.) of known operons, and subsequently these models are used as operon predictors (supervised methods). According to the differences in model generation, these supervised methods are summarized here.

1) One of the strongest operon predictors depends on the intergenic distances of adjacent genes, given the fact that genes within an operon tend to have much shorter intergenic distances than those at the borders of the operon. Based on experimental data on the intergenic distance of gene pairs within operons and at operon boundaries of the *E. coli* genome, a log likelihood function of intergenic distance for predicting operons is developed, and correctly identifies around 75% of the known *E. coli* operons [130].
2) The second method is to predict operons by detecting transcription control signals (e.g., existence of promoters and terminators). Construction of HMMs based on known promoters and terminators in *E. coli* enables the prediction of 60% of known operons [131].
3) The third method is based on the conservation of operon structures. Many sets of genes occur in conserved orders on multiple genomes across long stretches of evolutionary time, representing candidate operons.

A comparative genomics analysis on 34 prokaryotic genomes yield more than 7600 pairs of genes that are highly likely to belong to the same operon [132]. It also requires that adjacent genes in an operon are within a certain distance and that all genes in an operon are located on the same strand. This method allows highly confident prediction of operons in multiple species, but when it is applied to *E. coli*, a large portion of the known operons cannot be predicted [132]. The fairly low sensitivity of this method is due to the little conservation at the operon level between phylogenetically distant genomes [133].

4) The fourth method relies on the fact that genes in an operon tend to encode enzymes that catalyze successive reactions in metabolic pathways. The authors apply this method to 42 microbial genomes to identify putative operon structures, yielding a high prediction sensitivity as well as specificity [134]. This approach cannot make predictions at the whole-genome level since the information available does not span the whole genome.
5) The fifth method relies on the combined utilization of the above algorithms. Paredes et al. [135] present an operon map for the obligate anaerobe *Clostridium acetobutylicum* ATCC824 by combining intergenic distance, promoter prediction and rho-independent terminator prediction. Based on the set of known *C. acetobutylicum* operons, the presented operon map offers a prediction accuracy of 88%. Wang et al. [136] integrate several operon prediction methods, especially gene orientation analysis, intergenic distance analysis, conserved operon structure analysis and terminator detections, and develop a consensus approach to score the likelihood of each adjacent gene pair being in the same operon. Using this approach, a *Staphylococcus aureus* operon map is generated. When compared with a set of known *S. aureus* operons, this method successfully predict at least 91% of the gene pairs [136].

The efficiency of a supervised operon predictor depends largely on the type and amount of experimental information used for training. However, experimental information of operon structure is usually not available for a newly sequenced genome. Most of the existing operon predictors were originally built for *E. coli*, which has a large number of experimentally characterized operons. One may consider that these predictors are portable across genomes. Indeed, an operon predictor based on intergenic distances in *E. coli* [130] works fairly well when applied to *Bacillus subtilis* [137] and *Mycobacterium tuberculosis* [138]. The authors argue that the distance-based method has the possibility of operon prediction with high accuracy in most, if not all, prokaryotic genomes [137]. However, in many case operon predictors trained in a model organism are less portable when used for other target species, especially when these target organisms are phylogenetically distant from those used for training [139].

Because of the limitations of supervised algorithms when applied to genomes without extensive experimental investigations, unsupervised methods that do not require information about known operons for training have been developed recently for operon prediction [140–144]. Unsupervised methods for operon prediction are based on comparative genomic analysis of homologous genes across genomes. Supporting data for assignment gene pairs to an operon are collected from genomic sequence data and their functional annotations. These supporting data include intergenic distance, location on the same strand of DNA, conserved gene order, participation in the same metabolic pathway, similarity of protein functions, conserved gene functions across multiple genomes, promoter motifs, terminator signals and so on. An operon database, ODB, has been established using the unsupervised method to provide a data retrieval system not only of the known operons but also the putative operons predicted by the unsupervised methods [145]. At the time of publication this database contains information about 2000 known operons in more than 50 genomes, and about 13,000 putative operons in more than 200 genomes [145].

Prediction from microarray expression data

When microarray gene expression data are available, the accuracy of operon prediction is greatly elevated. For a simple two-sample experiment, a operon can be simply defined as a cluster of adjacent genes that have intergenic regions $< 50-100$ bp (different criteria used by different investigators [146, 147]) in length and are putatively transcribed in the same orientation and on the same strand, and that show the same tendency of up- or downregulation as determined by microarray. As a growing number of microarray gene expression experiments for a prokaryote become available, prediction of operons is practicable on the basis of co-expression patterns. Tjaden et al. [148] apply HMMs to estimate gene boundaries, which allows identification of 5′ untranslated regions of transcripts as well as genes that are operon members. A disadvantage of this method is that it uses a single source of microarray data. Bockhorst et al. [149] successfully predict operons by applying probabilistic language models to both DNA sequence and microarray expression data, which results in more accurate predictions than either alone. Both of these approaches use data from Affymetrix arrays that monitor expression of both coding and non-coding intergenic regions. However, the lack of intergenic probes in routine cDNA microarray experiments currently restricts the general application of these approaches.

Sabatti et al. [150] compiled data from 72 cDNA microarray experiments performed on *E. coli*, including compari-

sons of expression change between mutant and WT and studies in WT cells under different growth conditions. The correlation between expression ratios of adjacent genes across the microarray experiments was then used in a Bayesian classification scheme to predict whether the genes are in an operon or not, which allows a significant refinement of the sequenced-based predictions. Yamanishi et al. [151] applied a generalized kernel canonical correlation analysis to group genes, which share similarities with respect to position within the genome and gene expression. However, this method was restricted to subsets of *E. coli* genes that comprised known metabolic pathways. Bockhorst et al. [152] present a probabilistic machine-learning approach to predict operons using Bayesian networks. This approach exploits diverse evidence sources including gene coordinates, operon length, promoter and terminator signals, codon usage frequency and cDNA microarray expression data. Steinhauser et al. [153] propose a hypothesis-driven co-clustering strategy of genome sequence information and gene expression data that was designed to monitor occurrence of constitutive and conditional usage of transcription units in independent gene expression profiling experiments, allowing the identification of operons with high accuracy.

Verification by RT-PCR

The most credible situation is the experimental validation of operons by RT-PCR [154]. Given that genes in an operon are transcribed to a single RNA molecule, reverse transcriptase enzyme is used to synthesize first-strand cDNA that is subsequently used as a template for PCR amplification of products from the beginning, middle and end of a multi-gene cluster (Fig. 8), so as to define where the transcript from the multi-gene cluster starts and where it stops (see examples in [14]). RNA samples should be treated with DNase to avoid any contamination of genomic DNA. In some case, self-priming of the RNA, perhaps as a result of contamination of small RNA fragments, may provide a suitable 3′-terminus to prime the reverse transcriptase [155]. Experiments should be accompanied by subtle controls; (i) RNA but not RT primers, (ii) RNA but not reverse transcriptase, (iii) water as blank template, and (iv) purified genomic DNA were added respectively. Reactions (i), (ii) and (iii) must yield no detectable product.

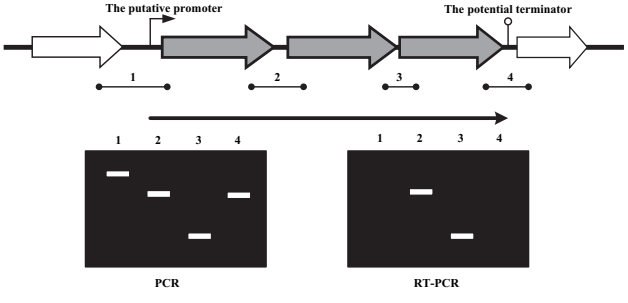

Figure 8. Verification of a putative operon by RT-PCR. Arrows represent the length and direction of transcription of the genes on the genome. The horizontal arrow depicts the putative primary transcript. The arrowheads indicate the location of primer pairs and amplicons. The cDNA and genomic DNA samples are analyzed by RT-PCR and PCR, respectively. PCR products are viewed with agarose gel electrophoresis.

Characterization of transcription factor-DNA interactions

ChIP-chip: mapping transcription factor binding sites on a genome

The chromatin immunoprecipitation (ChIP) assay has been historically used in conjunction with PCR to study protein-DNA interactions at a small number of specific DNA sites [156]. The recent adaptation of ChIP to DNA microarrays (chip) resulted in the method of 'ChIP-chip' (Fig. 9) for globally discovering genomic regions occupied by DNA-binding TFs in a living cell [157]. In ChIP-chip experiments, the nucleoprotein in the cells is cross-linked with formaldehyde, extracted and then sheared. Antibody against a TF of interest is then used to enrich the TF-cross-linked DNA fragments. The enriched DNA (referred to as 'IP DNA') is amplified by PCR and fluorescently labeled. As a control, sheared DNA from the formaldehyde cross-linking that has not been subjective to immunoprecipitation (referred to as 'control DNA') is similarly amplified and labeled with a different fluorescein dye. Finally, the differentially labeled DNA samples are mixed and co-hybridized to a microarray composed of DNA or oligonucleotide probes that represent the regions of the genome that one would like to probe for binding of the TF of interest. An enrichment factor is calculated that denotes the extent to which each genomic region is enriched by immunoprecipitation relative to the control DNA. ChIP-chip provides a genome-wide view of protein-DNA interactions with the mapping of TF-binding sites (TFBSs) on large swaths of the genome, giving a comprehensive understanding of where the TFs interact with the genome *in vivo*.

Cross-linking of DNA and proteins is required to fix the TF of interest to its binding sites. Formaldehyde is the most commonly used because the cross-links it forms are heat-reversible, permitting the downstream amplification of the immunoprecipitated DNA. Formaldehyde cross-links protein to both DNA and protein, and thus alternative cross-linking agents have been proposed [158]. The extent of cross-linking is critical and depends on the protein of interest. Cross-linking is generally carried out for a few minutes (5–20 min). Too much cross-linking may mask the epitopes of TFs, and too little cross-linking may lead to incomplete fixation. A time-course experiment is always performed to optimize cross-linking conditions.

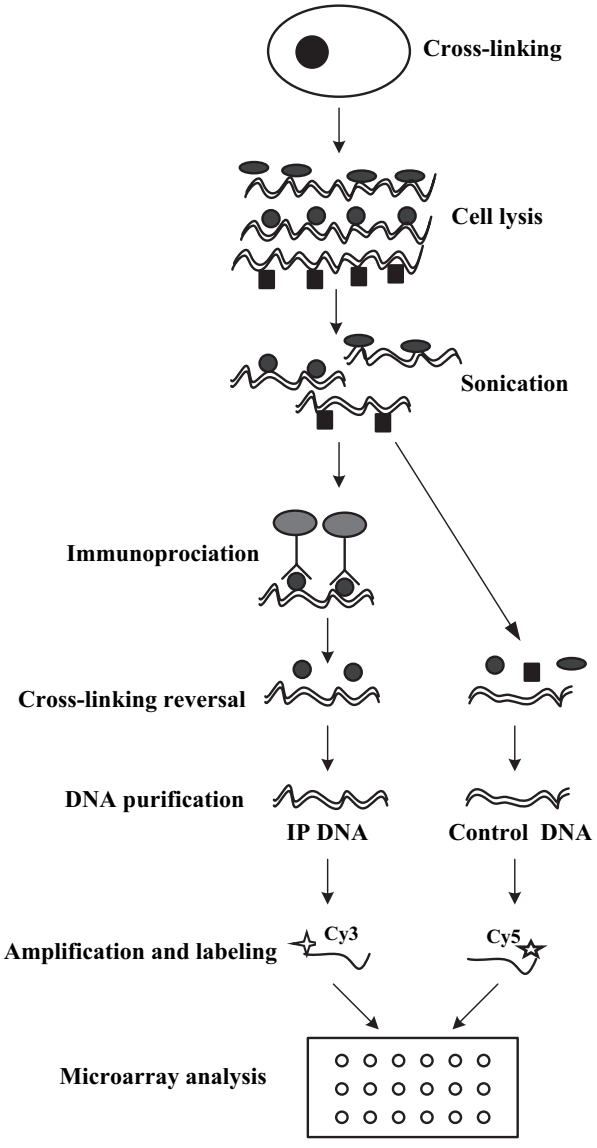

Figure 9. Procedures for ChIP-chip analysis. The nucleoprotein in the cells is cross-linked, extracted and then sonicated to give sheared DNA fragments. Antibody against a DNA-binding TF is then used to enrich the TF-cross-linked DNA fragments. The enriched DNA (IP DNA) and the sheared DNA from cross-linking that had not been subject to immunoprecipitation (control DNA) are amplified respectively, and labeled with different fluorescent dyes. The dual-fluorescently labeled DNAs co-hybridize to a microarray imprinted with the promoter DNA samples.

Fragmentation of the chromatin is required to make the TF-DNA interactions accessible to antibody for immunoprecipitation. Sonication is conducted using lower shearing power and turning the power on gradually. Samples should be kept on ice at all times to avoid denaturing of chromatin, as sonication generates heat. Micrococcal nuclease can also be used to digest chromatin [159], but sonication is generally preferred as it creates randomly sized DNA fragments, with no section of the genomic regions being preferentially cleaved by the micrococcal nuclease. Sheared chromatin DNA with length of 200–1500 bp (1–4 nucleosomes) can give a good resolution in mapping TF-DNA interactions. Optimal sonication conditions depend on cell type, cell concentration and sonicator equipment, including the power settings and number of pulses. In order to determine the ideal conditions for sonication, one should carry out a preliminary experiment where a cell lysate is sonicated for various time lengths, and the size of the DNA fragments is determined by agarose gel electrophoresis.

A well-characterized antibody is crucial in ChIP because it must specifically recognize its antigen fixed to chromatin DNA in solutions (see [158] for a method of determining the efficiency of an antibody to immunoprecipitate its target antigen). Antibodies for ChIP are ideally affinity-purified [160], but some investigators use antisera as an antibody source [161]. A polyclonal antibody is thought to be preferable to a monoclonal one, since the polyclonal antibody consists of a number of molecules that recognize different epitopes, which will reduce the probability of all epitopes being masked by cross-linking. Preliminary immunoprecipitation experiments should be performed to determine the appropriate amount of antibody to be used. Generally, 2–5 mg of antibody is used for every 20–50 mg of pure monosomes (a monosome is a complex of two subunits of the ribosome).

Low DNA yields (commonly 10–100 ng) from ChIP usually require DNA amplification, applied to both IP and control DNA samples, for downstream microarray detection. Randomly primed [162] or ligation-mediated PCR-based [163] methods have been most commonly used. Interestingly, use of microarrays containing oligonucleotide probes of large size (60 bp) increased the sensitivity greatly, allowing the authors to analyze less than 0.5-μg DNA samples, obtained directly from ChIP, without any amplification [164]. For fluorescent labeling, Cy5 or Cy3 conjugated nucleotide triphosphates can be directly incorporated into amplicons [157]. However, this may lead to labeling bias; for example, Cy5 tends to incorporate more readily than does Cy3. To reduce the influence of labeling bias, the incorporated dye is reversed in the dual-fluorescently labeled DNA samples for separated microarray hybridization (dye swap). Alternatively, indirect methods of labeling incorporate a non-fluorescent nucleotide analogue such as aminoallyl dUTP, followed by chemical conjugation of the cyanine-dye to the incorporated nucleotide analogue [165], which helps to eliminate the incorporation biases occurred in direct labeling.

Comparison of IP and control DNA samples by single-locus PCR is recommended after ChIP assay (Fig. 10). The creation of DNA amplicons for microarray detection cannot proceed unless the signal obtained in the control PCR shows a higher signal in the IP DNA sample than in the control DNA sample. In this approach, the PCR prim-

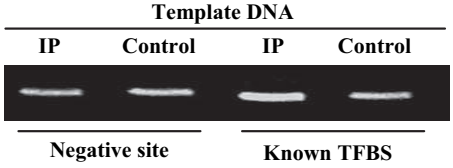

Figure 10. Single-locus PCR for quality control of ChIP. Primer pairs are designed from a known binding site of the TF of interest. For a successful ChIP assay, DNA fragments containing this binding site are enriched. Thus, the amount of PCR product using the IP DNA as template must be much higher than that using the control DNA as template. PCR amplifications targeting a known negative site for the TF are used as normalization.

ers are designed from one or two known binding sites of the TF of interest and a known negative site as the internal control, respectively.

In contrast to a large number of reports in yeast and human (reviewed in [166–168]), fewer ChIP-chip studies have been performed using prokaryotic genomes [161, 164, 169–173]. DNA microarrays used in these prokaryotic reports can be assigned to three types: microarrays mechanically spotted with PCR products [161, 169, 170, 172], Affymetrix arrays composed of oligonucleotides that are synthesized *in situ* [171] and high-density arrays spotted with oligonucleotides [164, 173]. All [169, 170, 172] or almost all [161, 171] of the probes represented in the microarrays correspond to coding sequences; these microarrays have traditionally been used for gene expression studies. Since the binding sites for TFs in prokaryotes generally lie relatively upstream of the coding regions for the genes that they control, the signals detected in these two kinds of microarrays may arise chiefly from the overlap of the fluorescently labeled probes with either the sheared DNA fragments with a TFBS or nearby coding sequence. These experiments might fail to identify some target sites and might identify a neighboring gene in addition to or even instead of the actual target [170]. The high-density arrays used by the investigators [164, 173] are spotted with oligonucleotides that space at regular intervals across a genome. The most robust microarray design for ChIP-chip is one having contiguously tiled DNA fragments that represent the entire genome (tiled microarrays) [168]; nevertheless, unwanted cross-hybridizations may occur. Microarrays spotted with PCR products (about 500 bp) or oligonucleotides (60 bp) corresponding to the upstream region of each annotated gene may be an alternative (promoter-specific microarrays). Promoter-specific microarrays are valuable in particular when TF-DNA binding is confined to cis-regulatory sequences close to coding regions, and thus they are very applicable in prokaryotes. Tiled microarrays are advantageous because they do not require prior knowledge of potential binding sites, and they allow one to utilize the 'neighbor effect' (see below) to precisely locate TFBSs [174].

ChIP-chip combined with microarray expression profiling

ChIP-chip assay and microarray expression experiments are complementary. Microarray-based regulon studies semi-quantitatively identify genes under either positive or negative control of a TF, but have difficulty distinguishing between direct and indirect targets. A TF, especially a global regulator, may indirectly control various cellular pathways by acting on other regulatory proteins. In addition, when a TF-encoding gene is deleted, some of target genes affected by the mutation may have other (not regulatory) secondary cellular affects.

ChIP-chip gives us a global understanding of where TFs interact with DNA, but in some cases genomic regions at which TF-binding is observed are not physiological sites at which TF stimulates or represses transcription *in vivo*. Several reasons [175–177] have been presumed into account for the occurrence of false positives ChIP-chip: (i) these sites are conditional cis-acting elements whose regulatory activity depends on other factors or unknown growth conditions; (ii) they serve as the storage sites of TFs; (iii) they are involved in the regulation of non-coding transcripts; (iv) there may be fortuitous binding sites with no function at all.

Combined analysis of transcriptome and ChIP-chip data will correlate the mapping of TFBSs with genes whose expression is dependent on a TF. Genes that are located at or near a site of TF binding as judged by ChIP-chip, and with transcription that is influenced by the disruption of TF as determined by microarray expression analysis, are most likely targets of direct regulation by the TF tested. This kind of incorporated analysis (see examples in [170]) provides a relatively small set of candidates that can be further tested by traditional biochemical methods, but it likely misses some of genes actually under the direct control of a TF. These missing genes may include (i) false negatives in ChIP-chip that result from failure to amplify some parts of the enriched chromosome DNA by PCR or low efficiency of formaldehyde cross-linking at some promoters [173], or (ii) genes identified by ChIP-chip are transcribed at a level too low to be detected by microarray expression experiments.

Computational promoter analysis combined with genome-wide screening experiments

Patterns, strings and matrices

In contrast to restriction enzymes that bind only to a unique and exactly defined DNA sequence, TFs recognize DNA sites containing variations and thus usually bind to multiple target sequences with varying affinity. This means the binding sites of a given TF on a prokaryotic genome also vary. However, most of the regulatory signals in these binding sites are carried in a short (5–

20 bp) and relatively conserved sub-region. This region represents the predominant contacts with the TF. If a collection of binding sites of a given TF have been defined from its target genes by DNase I footprinting, sequence alignments of these TFBSs will generate consensus patterns (Fig. 11). As an overrepresented motif recognized by a TF, a consensus pattern can be represented as either a consensus string or a position-specific scoring matrix (PSSM). A string is either a contiguous oligonucleotide (e.g. TAGTCGCACTA) or a dimer, W1NxW2, where W1 and W2 are short oligonucleotides separated by x arbitrary bases [178]. The bipartite characteristic represented by W1NxW2 results from the fact that many prokaryotic TFs have two DNA-binding regions, because of either the dimerization of the TF or the presence of two DNA-binding domains in a single protein. Thus, the cor-

(a) Position frequency matrix (PFM)

Nucleotide \ Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	64	61	51	11	12	5	10	96	22	26	26	61	33	49	6	14	97	22	76
T	49	42	46	88	12	111	11	7	57	25	47	21	26	31	100	8	14	13	18
C	7	10	13	18	7	7	2	15	27	47	35	19	26	22	10	104	1	88	10
G	8	15	18	11	97	5	105	10	22	30	20	27	43	26	12	2	16	5	24

(b) Consensus string

W W W T G T G A T C Y A R A T C A C A

(c) Position weight matrix (PWM)

A	0.7	0.6	0.5	-1.1	-1.0	-1.8	-1.1	1.1	-0.4	-0.2	-0.2	0.6	0.0	0.4	-1.6	-0.8	1.1	-0.4	0.9
T	0.4	0.3	0.4	1.0	-1.0	1.2	-1.1	-1.5	0.6	-0.2	0.4	-0.4	-0.2	0.0	1.1	-1.4	-0.8	-0.9	-0.6
C	-1.5	-1.1	-0.9	-0.6	-1.5	-1.5	-2.7	-0.7	-0.2	0.4	0.1	-0.5	-0.2	-0.4	-1.1	1.2	-3.3	1.0	-1.1
G	-1.4	-0.7	-0.6	-1.1	1.1	-1.8	1.2	-1.1	-0.4	-0.1	-0.5	-0.2	0.3	-0.2	-1.0	-2.7	-0.7	-1.8	-0.3

(d) Sequence-Logo representation

Figure 11. The position-specific scoring matrix of the *E. coli* CRP regulator. (*a*) A frequency matrix describes the alignment of binding sites the *E. coli* CRP regulator (see a review on the CRP regulator in [248]). The matrix contains $f_{b,i}$ that denotes the frequency of nucleotide *b* at position *i*. The data for this alignment consist of 128 known CRP-binding sites that are available in the RegulonDB database [77]. (*b*) The consensus string generated from the frequency matrix in (*a*) using the *convert matrix* tool, a part of RAST [183]. W, A or T. Y, C or T. R, A or G. The consensus string of *E. coli* CRP has been traditionally annotated as AAATGTGATCTAGATCACATTT or $TGTGAN_6TCACA$) [249]. (*c*) A weight matrix derived from the frequency matrix in (*a*) using the following formula [250]:

$$p(b, i) = \frac{f_{b,i} + s}{N + 4s}$$

$$W(b, i) = \log \frac{p(b, i)}{p(b)}$$

where $p(b, i)$ indicates the probability of nucleotide *b* at position *i*, *s* is the pseudocount used to replace zeros to avoid $\log(0)$, $W_{b,i}$ is the resulting weight and $p(b)$ is the background probability of nucleotide *b*. (*d*) The *sequence logo* [251] representation generated by the WebLogo tool [252].

responding two conserved regulatory motifs (generally each with a length less than 10 bp) can be found in the TFBSs [179–181]. A major drawback of the consensus strings is that they remove much of the information originally present in the set of TFBSs. In contrast, a PSSM retains most of the information and is better suited to evaluate new potential sites [182]. In the PSSM, each row represents a position and each column a nucleotide. Representation of consensus patterns with PSSMs can give a full description of the uneven composition in each position, i.e. some nucleotides occur much more frequently than others.

Pattern discovery

Computational promoter analysis serves to predict the consensus pattern de novo from a set of DNA sequences revealed by either ChIP-chip or microarray expression analysis (pattern discovery). Microarray expression experiments can reveal wide sets of genes whose transcription is affected by an environmental perturbation or the disruption of a TF of interest. Upstream promoter-proximate sequences of these differentially expressed genes can be retrieved from the genomes using specific tools, for instance, *retrieve-seq* [183]. Ideally, these differentially expressed genes are assigned into various putative operons (see above) before the collection of promoter sequences; in this situation, the upstream sequence of every first gene in each operon is subsequently collected. Despite the conserved motifs recognized by a given TF being represented only by small DNA fragments rather than the large surrounding sequences, and the indirect targets of the TF being mixed with the direct targets, searching and compilation of potential motifs in the promoter sequences with specific pattern discovery algorithms will build regulatory patterns from an array of differentially expressed genes [184] or a specific cluster of co-expressed genes [185].

Dozens of pattern discovery algorithms have been developed in the past few years (Table 2). Systems that integrate versatile tools are also available [183, 186]. Here, we give an example of a mix mode proposed by Conlon et al. [187] to discover regulatory motifs, for it works well for microarray mutant expression data from both a single two-sample experiment and multiple time-course measurements. In their approach, MDscan [188] is first used to generate a large set of non-redundant candidate motifs that are enriched in the DNA sequence upstream of genes with the highest-fold change in mRNA level, under the assumption that genes with the most dramatic increase or decrease in mRNA expression are most likely to be directly regulated by the TF, and that these might contain strong TFBSs. Motif Regressor [187] then scans the promoter region of every gene in the genome with each candidate motif to measure how well a promoter matches a motif (in terms of both number of sites and strength of matching). It then uses linear regression analysis to select motifs whose promoter-matching scores are significantly correlated with downstream gene expression values. When ranking motifs by linear regression *p*-value, Motif Regressor automatically picks the best motif and optimal motif width.

ChIP-chip can map the probable TF-DNA interaction loci within 1–2-kb resolution. Depending on the efficiency of chromatin fragmentation and the resolution of the arrayed DNA elements, arrayed probes representing genomic regions both at the binding site and near the binding site may be detected as ChIP-enriched elements. In addition to this neighbor effect, noise may come from the inherent false positives observed in ChIP-chip [173]. That notwithstanding, the ChIP-chip data provide much more accurate information about the genome-wide location of *in vivo* TF-DNA interactions compared with the microarray expression data. Investigators have developed various computational methods [188–191] that can examine selected ChIP-chip sequences and search for DNA sequence motifs over-representing the TF-DNA interaction sites (Table 3).

In spite of the abundance of existing tools for pattern discovery, most of them provide little information for further evaluation. Current pattern discovery algorithms are far from perfect. Hu et al. [192] designed a comprehensive set of performance measures and benchmarked five modern sequence-based motif discovery algorithms using large datasets generated from the RegulonDB database [77]. Several factors have been shown to affect prediction accuracy, scalability and reliability. Limitations of these algorithms come from the inherently low signal/noise ratio in purely sequence-based motif discovery problems, the pattern model used to capture regularity among the TFBSs and finally local optima phenomena in optimization algorithms. However, the authors argue the potential of improvement in these algorithms and suggest several promising directions for further improvements. In addition, Tompa et al. [193] described an assessment of 13 different computational tools for de novo prediction of regulatory elements, using eukaryotic data sets derived from the TRANSFAC database [194] and found that the absolute measures of correctness of these programs are low.

Pattern matching

When the consensus pattern for a given TF is either known from the literature or databases, or generated as described above, one may subsequently find homologues of these DNA patterns in the upstream sequences of a set of genes from ChIP-chip or microarray expression experiments (pattern matching), and even scan the whole genome (whole-genome pattern matching) to predict candidate target genes [195]. These computational approaches (Table 2) provide a systematic test for determining whether a gene is

Table 2. Selected tools for pattern discovery and pattern patching.

Program	Pattern	Description	Reference	URL
Pattern discovery in upstream regulatory regions of co-expressed genes				
MEME	matrix	It fits a mixture model by expectation maximization to discover motifs.	[255]	http://meme.nbcr.net/beta/
CONSENSUS	matrix	Uses a greedy algorithm searching for the motifs with maximum information content.	[256]	http:/bifrost.wustl.edu/consensus/
Gibbs Motif Sampler	matrix	The original Gibbs sampling strategy.	[257]	http://bayesweb.wadsworth.org/gibbs/gibbs.html
AlignACE	matrix	It judges alignments sampled during the course of Gibbs sampling with a maximum a priori log likelihood score, which gauges the degree of over-representation.	[258, 259]	http://atlas.med.harvard.edu/
MotifSampler	matrix	Extends Gibbs sampling for motif finding with a higher-order background model.	[260, 261]	http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html
BioProspector	string/matrix	Finds motifs using a Gibbs sampling strategy modified with the zero to third-order Markov background models, followed by judgment with a Monte Carlo method.	[262]	http://ai.stanford.edu/~xsliu/BioProspector/
Weeder Web	matrix	The Weeder algorithm finds all patterns P that occur in at least q sequences of a set, with at most $e = \|\varepsilon\|P\|\|$ mutations.	[263]	http://159.149.109.16:8080/weederWeb/
Motif Regressor	matrix	Uses linear regression analysis to select the best motifs by ranking *p*-value. MotifRegressor relies in part on MDScan.	[187]	http://www.techtransfer.harvard.edu/Software/MotifRegressor/
CisModule	matrix	Based on the hierarchical mixture model, CisModule is developed for the Bayesian inference of module locations and within-module motif sites.	[264]	http://www.people.fas.harvard.edu/~qingzhou/CisModule/
Pattern discovery applied to ChIP-chip data				
MDscan	string/matrix	It selects several top motif candidates according to the chip-chip enrichment ratios to build motif models and then employs a greedy strategy to improve the models.	[188]	http://ai.stanford.edu/~xsliu/MDscan/
MotifBooster	matrix	A boosting approach to modeling TF-DNA binding.	[189]	http://biogibbs.stanford.edu/~hong2004/MotifBooster/

Table 2. (Continued).

Program	Pattern	Description	Reference	URL
Pattern matching for user-defined DNA sequences				
FUZZNUC	string	FUZZNUC uses prosite style patterns to search nucleotide sequences. It intelligently selects the optimum-searching algorithm to use, depending on the complexity of the search pattern specified.	[265]	http://bioweb.pasteur.fr/seqanal/interfaces/fuzznuc.html
MAST	matrix	It utilizes the product of *p*-values scoring method to valuate the scores for matching a sequence to a motif. It allows for predictive screens of entire genomes.	[266]	http://meme.sdsc.edu/meme/mast-intro.html
MatInspector	matrix	Includes position weighting of the matrices based on the information content of individual positions and calculates a relative matrix similarity.	[267]	http://anthea.gsf.de/biodv/matinspector.html
Match	matrix	It uses two score values, the matrix similarity score and the core similarity score.	[268]	http://www.gene-regulation.com/pub/programs.html#match
P-Match	matrix	The P-Match search algorithm utilizes PWMs from TRANSFAC and computes a *d*-score value that measures the similarity between a sub-sequence of the length in DNA and a given TF site from the site set V.	[269]	http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi
PatSearch	matrix	It is a flexible and fast pattern matcher able to search for specific combinations of oligonucleotide consensus sequences, secondary structure elements and position-weight matrices.	[270]	http://www.ba.itb.cnr.it/BIG/PatSearch/
Whole-genome pattern matching				
The *coli*BASE pattern search	string	It allows for searching within bacterial genomes for short DNA sequence patterns using FUZZNUC.	[271]	http://colibase.bham.ac.uk/pattern/index.cgi?help=pattern&%20frame=pattern
MSCAN	matrix	MSCAN evaluates the combined statistical significance of sets of potential TF binding sites. When the significance level, or *p*-value, falls below a chosen threshold value, a prediction is output.	[272]	http://mscan.cgb.ki.se/cgi-bin/MSCAN
MotifViz	matrix	The MotifViz web server contains four programs, Possum, Clover, Rover and Motifish.	[273]	http://biowulf.bu.edu/MotifViz
Displaying consensus patterns				
WebLogo	Web-based tools applied to generate sequence logos graphically displaying consensus sequences.		[252]	http://weblogo.berkeley.edu/
enoLOGOS			[274]	http://biodev.hgen.pitt.edu/enologos/
Tool suites				
RSAT	Web-based tool suites dedicated to predict regulatory sites in non-coding DNA sequences, including sequence retrieval, pattern discovery, pattern matching, whole-genome pattern matching and other utilities.		[183]	http://rsat.ulb.ac.be/rsat/
TOUCAN			[186]	http://homes.esat.kuleuven.be/~saerts/software/
INCLUSive			[275]	http://aulne8.esat.kuleuven.be:8080/inclusive/index.jsp

likely under the direct control of a given TF, and a framework for continued biochemical analysis.

A big problem of these matching approaches is the fairly large number of false positives. Given the short size of the consensus sequences and the large size of the input sequences, especially complete genomes, a large array of matches could be returned after a simple running. Combining a set of functionally related TFs [196] and searching for their co-abundance [197] can significantly increase specificity [198].

Another drawback is that the consensus patterns used largely limit the computational searches. The most reliable patterns used for searching come from the alignment of the available binding sites determined by DNase I foot-

Table 3. Public databases for prokaryotic transcriptional regulation.

Database	Features	Reference	URL
DDBJ	all known nucleotide and protein sequences; for some genes, there is information on location of transcription start point and TFBSs	[276]	http://www.ddbj.nig.ac.jp
EMBL		[277]	http://www.ebi.ac.uk/embl.html
GenBank		[278]	http://www.ncbi.nlm.nih.gov/Entrez
ArrayExpress	microarray gene expression data and online analysis tools	[279]	http://www.ebi.ac.uk/arrayexpress
SMD	microarray data along with many tools to explore and analyze those data	[280]	http://genome-www.stanford.edu/microarray
DBD	predicted transcription factor repertoires for 150 completely sequenced genomes, their domain assignments and the hand-curated list of DNA-binding domain HMMs	[23]	http://stash.mrc-lmb.cam.ac.uk/skk/Cell2/index.cgi?Home
Extra-TRAIN	extragenic regions and transcriptional regulators of 230 genomes of bacteria and archaea		http://www.era7.com/ExtraTrain
BacTregulators	transcriptional regulators of AraC and TetR families	[281]	http://www.bactregulators.org
PRODORIC	detailed information about operon and promoter structures, including huge collections of transcription factor binding sites	[282]	http://prodoric.tu-bs.de
ODB	Information about 2000 known operons in more than 50 genomes and about 13,000 putative operons in more than 200 genomes	[145]	http://odb.kuicr.kyoto-u.ac.jp
TRACTOR_DB	predicted new members of 74 regulons in 17 gamma-proteobacterial genomes	[283]	http://www.tractor.lncc.br
BIND	biomolecular interaction network database that contains complete information about interactions and reactions arising from biopolymers (protein, RNA and DNA), as well as small molecules, lipids and carbohydrates.	[284]	http://www.bind.ca
DBTBS	*Bacillus subtilis* promoters and TFs	[285]	http://dbtbs.hgc.jp
MtbRegList	regulatory DNA motifs, TFs and experimentally identified transcription start points in *Mycobacterium tuberculosis*	[286]	http://www.USherbrooke.ca/vers/MtbRegList
EcoCyc	a comprehensive source of information on promoters, operons, genetic networks, TFBSs, functionally related genes, protein complexes and protein-ligand interactions in *E. coli*	[224]	http://ecocyc.org
RegulonDB	promoters, TFs, TFBSs, terminators, operons, regulons, transcriptional regulatory networks, and growth conditions in *E. coli*	[77]	http://regulondb.ccg.unam.mx/index.html
DPInteract	binding sites for *E. coli* DNA-binding proteins	[287]	http://arep.med.harvard.edu/dpinteract
PromEC	*E. coli* promoters with experimentally identified transcriptional start sites	[288]	http://margalit.huji.ac.il/promec

printing, but in many cases the collection of these binding sites is too small for sufficient coverage. When the authors tested for OxyR binding to six targets predicted to have high scores in a computational search in *E. coli* with a motif based on nine known OxyR binding sites, only three of them were found to be bound by OxyR in DNase I footprinting assays, whereas one predicted binding site with a low score was revealed to be bound with high affinity [199].

An additional problem in pattern searching is that some TFs have no consensus sequence common to all or almost all of their target genes. For example, the PhoP regulator has conserved Mg^{2+}-responsive modulation of gene expression [200], but the previously characterized (T/G)GTTTA(A/T) motif cannot be detected in many promoters newly discovered to be the direct targets of PhoP [201].

Biochemical dissecting of transcription factor-DNA interactions

Detection of direct binding of transcription factor to target DNA

As described above, EMSA has been used widely in detecting and verifying the direct association of candidate DNA fragments with a known sequence-specific DNA-binding protein [202–204]. Three controls in EMSA can be utilized to ensure the specificity of the TF-DNA interaction: (i) the most common test is to add unlabeled competitor DNA, including target DNA and non-target poly(dI-dC)·poly(dI-dC), to compete for the TF of interest. Specificity of binding is indicated when excess unlabeled target DNA reduces the amount of labeled TF-DNA complex, while excess non-target DNA has no effect [203, 204]; (ii) site-specific mutagenesis of the presumed DNA binding site can be used to examine specificity. Altering conserved nucleotides in the putative binding region may abolish TF-DNA interactions [205]; and (iii) another test of specificity is the 'supershift' assay. Antibody to the TF of interest added to the preformed TF-DNA complex can further retard its mobility (supershift) during electrophoresis [206].

Location of transcription factor binding sites

A DNase I footprinting assay is used to identify a precise TFBS at single-base pair resolution [207]. The end-labeled DNA probe incubated with the TF of interest is treated lightly with the restriction enzyme DNase I, which digests nucleic acids starting within the strand and makes single-strand breaks (nicks) in the DNA without damaging the bases (Fig. 12). With this mild digestion, some DNA molecules are not cut at all, and most are cut only once. Different molecules are cut in different places, so that one gets a family of labeled fragments ending at positions throughout the DNA. However, the DNA site bound by the TF is protected against restriction enzyme cleavage. From the position of the cleavage sites absent, the position and extension of the binding site can be deduced (see examples in [208]).

Determination of transcription start points

Primer extension can be used to map the 5′ terminus of an RNA transcript, which allows one to determine the start site of transcription and helps to localize the core promoter region [209]. The length of the cDNA reflects the number of bases between the labeled nucleotide of the primer and the 5′ end of the RNA, and the yield of primer extension product reflects the abundance of targeted RNA (Fig. 13) (see examples in [210]).

The above three methods are the most commonly used over the last 20 years for biochemical characterization of specific TF-DNA interactions. Reports using these methods can be found in almost every issue of high-quality microbial journals. As complementary experiments for verification of candidate TF targets that are identified through genome-wide screening methods, including ChIP-chip [170], microarray expression analysis [211] and computational prediction [199], they are now proving their greater utility in gene regulation research. In addition to prototypes using radiochemicals, non-radioactive derivatives have also been established [212–214]. A big advantage of these methods over traditional radioactive methods is that the DNA probe can be labeled with different fluorescein dyes, which provides simultaneous detection with capillary electrophoresis and automated DNA sequencing [215, 216].

Public databases for prokaryotic transcriptional regulation

The amount of both experimentally validated and computationally predicted knowledge of prokaryotic transcriptional regulation is ever increasing, providing important insights into a variety of biological processes. To make maximum use of these data, electronic databases have been widely developed in the past few years (Table 3). Most of them are integrated with useful tools that are either Web based or downloadable, as well as links to related Web sites and even training courses. These databases serve the scientific community as a repository for data to facilitate access and to be used subsequently for specific investigations. Given the attraction of unceasing improvement and easy access, more and more people in the community now appreciate the importance of databases in spreading knowledge. It should be noted here that the majority of database authors and curators receive little or no remuneration for their efforts and that it is still difficult to obtain money for creating and maintaining a database [217].

Four (EcoCyc, RegulonDB, DPInteract and PromEC) of the 14 databases listed in Table 3 are specific for *E. coli*, which represents the best-studied biological model and the primary reference organism. In addition to a long history of intense biochemical and genetic investigations, much research on computational biology, including transcription and regulation, has been reported on *E. coli* over the past few years. The large amount of accumulated knowledge on this bacterium constitutes the foundation for the proposal of the International *E. coli* Alliance (IECA) [209], and also strongly benefits current studies in genetics, genomics, transcriptomics, proteomics, bioinformatics and systems biology of every other organism. The four databases of *E. coli* represent the relevant knowledge in a computable and easy-to-use manner, providing a blueprint for predicting regulatory elements (promoters, TFs, TFBSs and operons), reconstructing the metabolic pathways with regulatory information and finally modeling regulatory networks.

From specific gene regulation to regulatory network

Network motifs

Transcriptional regulatory networks (TRNs), which control gene expression temporally in a cell, provide the solid framework for structural and functional analysis of gene regulation in an organism. The most basic components in TRNs are TFs and their target genes. The regulatory interactions – binding of TFs to the promoters of their target genes – in a TRN are usually depicted as a directed graph in which nodes are connected by edges [218]. Nodes re-

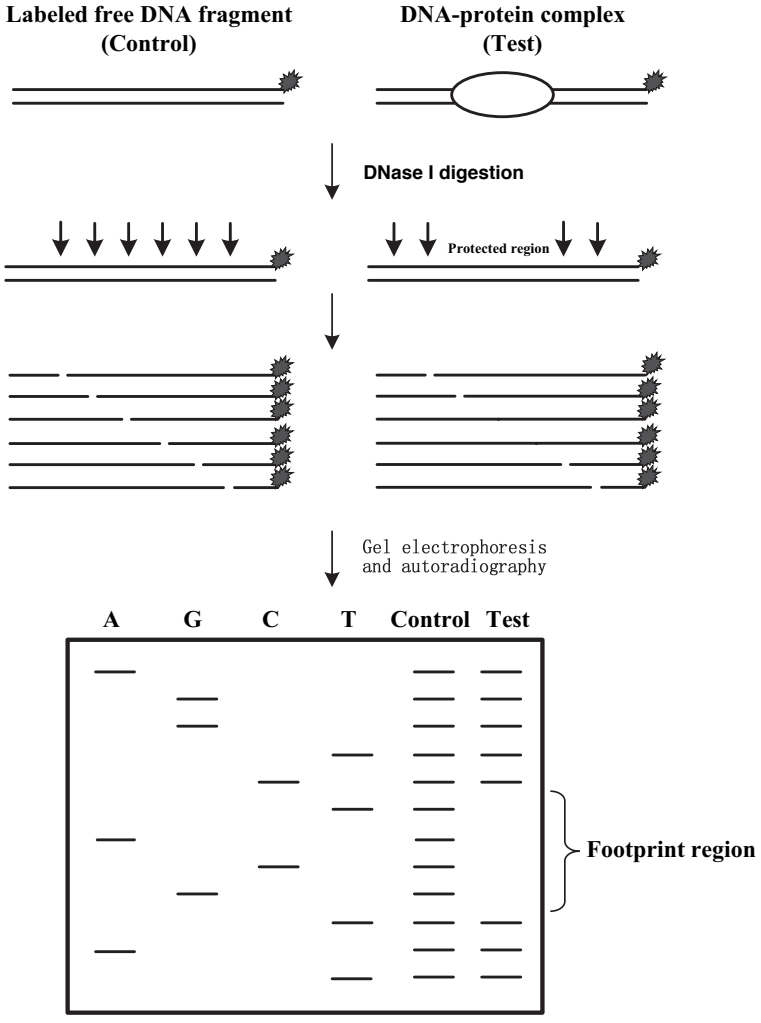

Figure 12. DNase I footprinting. Promoter DNA samples are generated by PCR. The noncoding or coding strand of promoter DNA is radioactively labeled, and incubated with a purified TF protein. After partial digestion with DNase I, the resulting fragments are analyzed by denaturing gel electrophoresis. The sequence ladders containing the products of a sequencing reaction are generated with the same primers used to synthesize the DNA fragment for DNase I treatment. The DNA sequence ladders are used as co-ordinates of the region protected against DNase I cleavage.

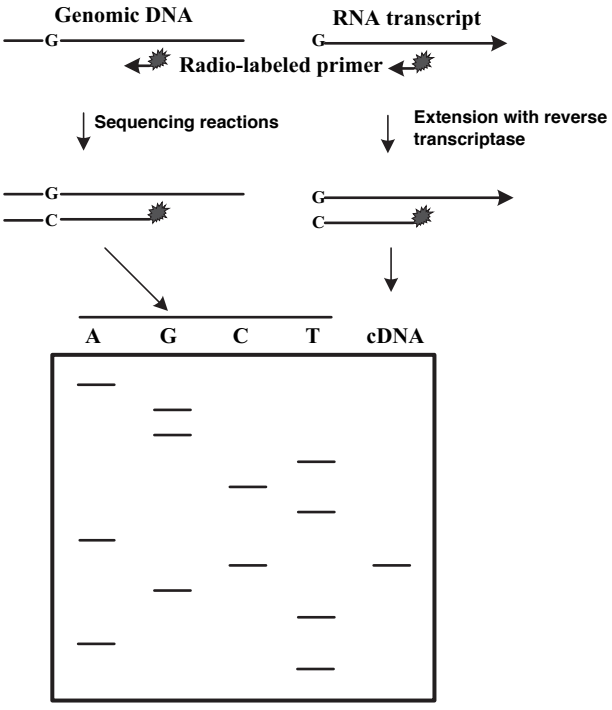

Figure 13. Primer extension. An oligonucleotide primer is designed to be complementary to a portion of the RNA transcript of each operon. The primer is end-labeled, hybridized to the RNA and extended by reverse transcriptase using unlabeled deoxynucleotides to form a single-stranded DNA complementary to the template RNA. The resultant cDNA is analyzed on a sequencing gel as for DNase I footprinting. To serve as sequence ladders, sequencing reactions were also performed with the same primers used for primer extension.

present TFs and their target genes, while edges represent direct regulatory interactions, either activation or repression.

Network motifs are defined as the over-represented patterns of topological interaction between nodes (TFs and target genes); they recur in many different parts of a network at frequencies much higher than those found in randomized networks [219, 220]. In general, the known true TRNs in bacteria and yeast can be categorized by six basic motifs (Fig. 14) [220, 221]. The first motif is the feed-forward loop in which the first TF regulates a second one and both regulate a common target gene. The second motif, called 'bin-fan', consists of two input TFs that bind together to two genes. The above two motifs appear to be the major network motifs found in bacteria and yeast. In contrast, the following four motifs are relatively rare in the existing TRNs [220, 221]: (i) a single-input module that is defined by a set of target genes that are controlled by a single TF; (ii) an autoregulation loop that consists of a regulator that targets itself; (iii) for a multi-component loop, two TFs that regulate each other; and (iv) in a regulator chain motif, a set of TFs that regulate one by one to constitute a regulatory chain.

Network motifs represent the simplest units of the network architecture, allowing an easily interpretable view of the TRNs [220]. Each of these motifs plays a specific information-processing role in the network. Network motifs can self-organize to produce TRNs because of the large ratio of genes to TFs in the genomes; in this way links that are already present in the motifs, without the addition of extra connections, define an extensive network that includes the majority of nodes in the entire network [222]. The stability of the TRNs to small perturbations is highly correlated with the relative abundance of these network motifs, which is a driving force defining the non-random organization of the networks [223]. It has been shown that TFs whose transcripts have short half-lives are significantly enriched in motifs [222]. This enrichment enables the network to adapt quickly to environmental changes and mitigates gene expression fluctuations, or internal noise.

The true transcriptional regulatory network in *E. coli*

The RegulonDB [77] and EcoCyc [224] databases contain a comprehensive set of experimental evidence on the direct regulatory interactions between TFs and their targets genes in *E. coli*, providing a prerequisite for the construction of the genome-wide true TRN. Although the earlier versions of these two databases have different content due to the variable use of gene names and synonyms, they are synchronized beginning with version 9.0 of EcoCyc and 4.4 of RegulonDB [225]. The TF-DNA interaction datasets in RegulonDB and EcoCyc was used

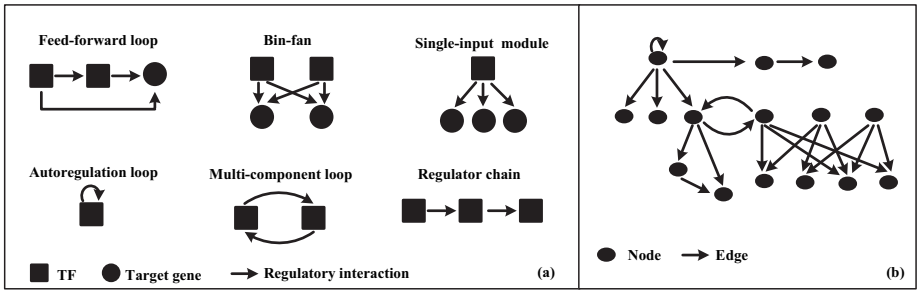

Figure 14. The six basic network motifs detected in the TRNs. (*a*) The network motifs. (*b*) A presumed TRN in which the six motifs in (*a*) can be found.

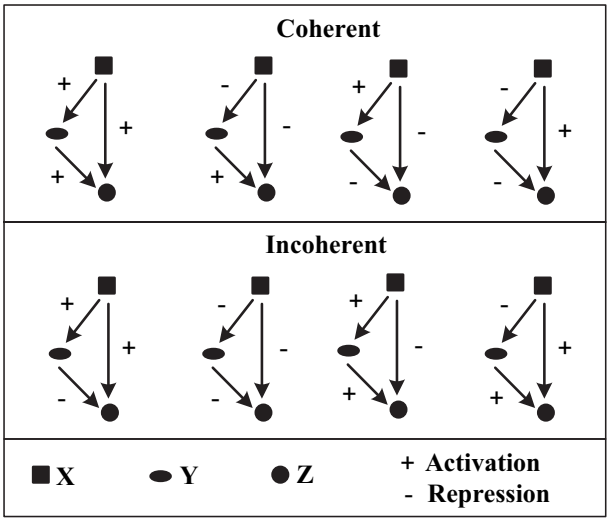

Figure 15. The eight types of FFLs. In the FFL, two TFs (X and Y) jointly regulate a single target gene (Z), meanwhile X controls Y. The FFL has three regulatory interactions, each of which can be either positive (activation) or negative (repression). Thus, there are in total eight structural types of these positive and negative interactions, four of which are termed 'coherent' – the sign of the direct regulation path (from X to Z) is the same as the overall sign of the indirect regulation path (from X through Y to Z) [220, 227, 228]. The other four types are called 'incoherent', for which the signs of the direct and indirect regulation paths are opposite. Some FFL types appear in the network more frequently than others [220, 227, 228].

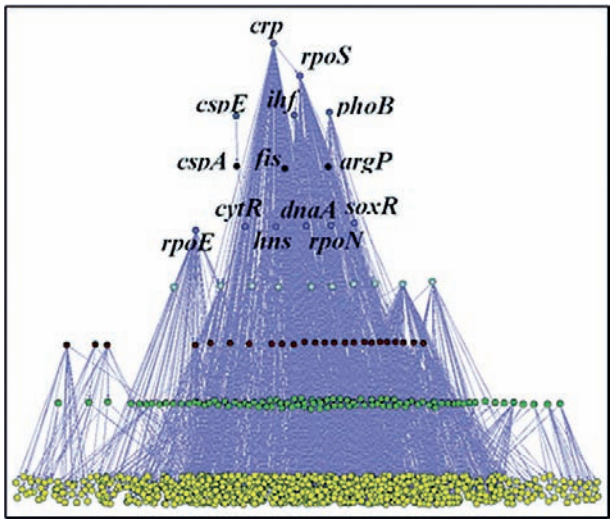

Figure 16. The hierarchical structure of the *E. coli* TRN. The *E. coli* TRN includes 1278 genes (rings) and 2724 interactions (arrows) [227]. There are nine layers in the regulatory hierarchy. Autoregulatory loops are not shown in this TRN.

to generate the genome-wide TRN of *E. coli* as early as 4 years ago, with an emphasis on identifying statistically over-represented motifs [220]. More recently, a extended *E. coli* TRN [226, 227] was reconstructed from RegulonDB and Ecocyc, with an emphasis on determining global topological properties.

The feed-forward loop (FFL) is the only three-node motif and the most predominant motif in the true TRNs of *E. coli* [220, 226, 227]. Theoretical analysis of the functions of the eight structural types of FFLs, as shown in Figure 15, indicates that the four incoherent FFLs act as sign-sensitive accelerators – they speed up the response time of the target gene expression following stimulus steps in one direction (e.g. off to on) but not in the other direction (on to off) – while the other four coherent FFLs act as sign-sensitive delays [228]. Thus, FFLs have important functions in controlling the dynamic response of the target gene. Both coherent and incoherent FFL behavior is sign sensitive; they accelerate or delay responses to stimulus steps, but only in one direction.

The newly defined genome-wide TRN of *E. coli* exhibits a distinct multi-layer hierarchical structure [227] (Fig. 16). Its primary features are the following:

1) Through the identification of a few multi-component loop (MCL) motifs in the network, further survey assigns the two genes in each MCL to a single operon and thus the same layer. The resulting straightforward top-down relationships in the TRN strongly indicate the lack of feedback regulation at transcription level. It is thought that feedback control might be through other interactions at the post-transcriptional level, rather than through TF-DNA interaction at the transcriptional level.
2) All of the known six network motifs can be detected in the network, while the three-node motif of FFL is most highly representative.
3) The distribution of the eight types of FFLs (see above) is different from that observed in the previous network [220]. In addition, in contrast to the previous notion that most motifs overlap and generate distinct homologous motif clusters and then clusters of different motifs are connected to make super clusters [229], most FFLs interact and form a giant motif cluster. Therefore, using a more complete and reliable network is important for investigating the structure and function of gene regulation.
4) The majority of genes are regulated by two or more interacting FFLs or other more complicated network motifs together with TFs not belonging to any network motifs. Only a small portion of the genes are solely regulated by only one FFL.
5) TFs within more top layers regulate many genes. Indeed, the previously identified global transcription regulators [10, 21] are located in the few topmost layers.

Modeling transcriptional regulatory networks from various sources of data

Microarray expression data represent the most widely available data source for the inference of TRNs. In par-

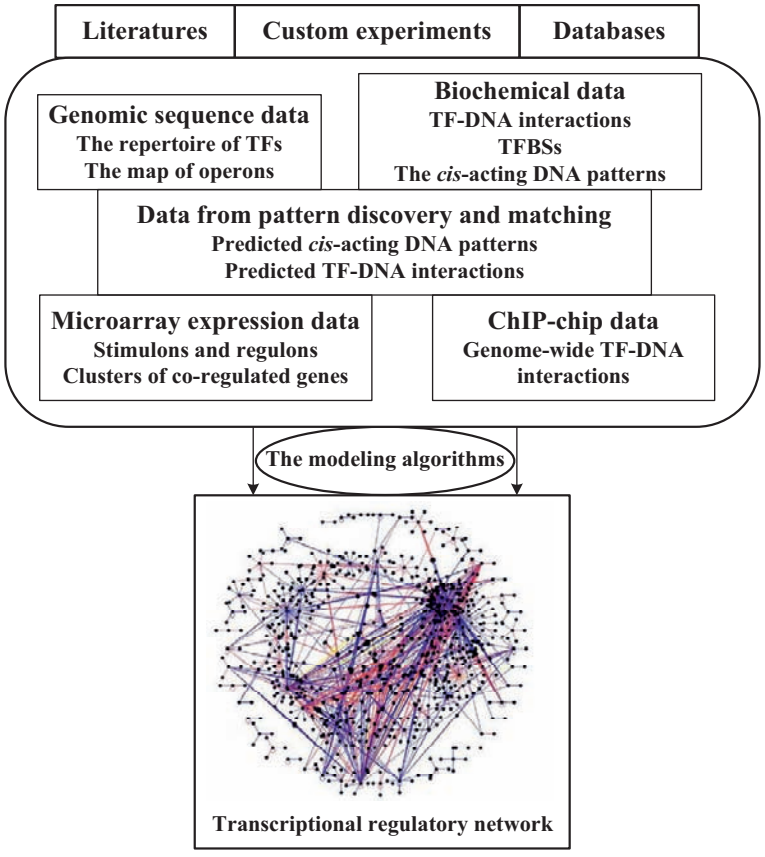

Figure 17. Modeling TRN from a combined source of data. Shown are various data sources for TRN modeling. A combination of these data will provide a much more sophisticated view of how individual genes are ranked in the TRNs.

ticular, genome-wide analysis of changes in gene expression in response to the disruptions of TFs produce wide sets of potential target genes for many TFs in a organism such as *E. coli* [230]. A common subsequent practice is to search for cis-acting DNA patterns in the upstream sequences of co-expressed genes revealed by microarray experiments, although it is often prone to inherent noise. Use of large-scale microarray expression data alone or in combination with computational promoter analysis has provided a powerful framework for TRN reconstruction [231–233]. Modeling TRNs in this context is far beyond the clustering analysis that only tells us which genes are co-regulated rather than what regulates what [234]. TF-binding data measured by ChIP-chip outline the ability of TFs to bind all regulatory regions on a genome, leading to great improvement in reconstructing the TRN structures over gene expression data [221]. However, current ChIP-chip studies on the prokaryotes only beginning. That notwithstanding, a combination of all these data will provide a much more sophisticated view of how individual genes are ranked in the TRNs (Fig. 17).

A variety of mathematical models have been applied to interfer genetic networks, including Boolean networks [235], linear models [236], Bayesian networks [237] etc. Several excellent reviews [238–240] address these issues that thus will not be discussed in this paper. Alternatively, statistical methods [241–243] have been proposed to identify modules of co-regulated genes from microarray expression data and/or ChIP-chip data. These methods can be divided into steps that first group genes into modules that are defined as genes co-regulated by one or more TFs, then relate each module to the cellular conditions or environmental stimuli that control it and finally discover connections between these modules to reconstruct the TRNs. Advances in compiling the interactions between TFs and target genes for the reverse engineering of TRNs will require the development of new and more powerful computational and visualization tools, especially those integrating diverse data types and transforming them into biological models. Algorithms are certainly proposed by experts in biostatistics, but the tools should be presented in a user-friendly format to allow numerous biological researchers to gain more information from their experiments.

Conclusions

Current efforts to measure global changes in gene expression with DNA microarrays, map genome-wide TF-

DNA interactions with ChIP-chip, find cis-acting DNA elements in the promoters of genes of interest by computational methods, and detect specific TF-DNA interactions and locate TFBSs within upstream sequences of the regulated genes with conventional biochemical techniques have already produced good understanding of the genetic circuitry of transcription regulation in prokaryotes. Continuing studies should identify more and more target genes of more and more TFs in prokaryotes, especially model organisms such as *E. coli*. This would provide needed data for reconstructing regulatory networks. A gene in cells may be regulated by different TFs, and the contribution from different TFs may function under different conditions. The relationships between TFs and structural genes may be much more complex than we imagine. A considerable challenge is thus to find novel environmental cues under which TFs trigger gene regulation [244]. Data from mRNA expression and TF-DNA interactions give only limited information that does not include post-transcriptional events and protein-protein or protein-metabolite interactions. The TRNs thus give only part of the picture of cell cycles. A complete genetic network should be a three-dimensional architecture involving regulators, enzymes, structural genes, functional RNAs and metabolites, which controls temporal changes in gene expression for growth, proliferation, adaptation and development. The genetic networks reconstructed in the future will be no doubt very complex. 'The more complex the networks become, the closer they are to mirroring the dynamic changes that occur in a living cell' [245].

Acknowledgements. We apologize to colleagues whose research is omitted here. We thank members of our laboratory for helpful comments and suggestions. Work in our laboratory has been supported by grants from the National High Technology Research and Development Program of China (program 863, nos. 2001AA223061 and 2004AA223110), the National Natural Science Foundation of China (nos. 30371284, 30471554 and 30430620) and the National Science Foundation of China for Distinguished Young Scholars (no. 30525025).

1 Borukhov, S. and Nudler, E. (2003) RNA polymerase holoenzyme: structure, function and biological implications. Curr. Opin. Microbiol. 6, 93–100.
2 Wosten, M. M. (1998) Eubacterial sigma-factors. FEMS Microbiol. Rev. 22, 127–150.
3 Browning, D. F. and Busby, S. J. (2004) The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2, 57–65.
4 Gralla, J. D. (1996) Activation and repression of *E. coli* promoters. Curr. Opin. Genet. Dev. 6, 526–530.
5 Barnard, A., Wolfe, A. and Busby, S. (2004) Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. Curr. Opin. Microbiol. 7, 102–108.
6 Marmorstein, R. and Fitzgerald, M. X. (2003) Modulation of DNA-binding domains for sequence-specific DNA recognition. Gene 304, 1–12.
7 Pabo, C. O. and Sauer, R. T. (1992) Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. 61, 1053–1095.
8 Huffman, J. L. and Brennan, R. G. (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. Curr. Opin. Struct. Biol. 12, 98–106.
9 Perez-Rueda, E. and Collado-Vides, J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. Nucleic Acids Res. 28, 1838–1847.
10 Martinez-Antonio, A. and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol. 6, 482–489.
11 Cotter, P. A. and Miller, J. F. (1998) *In vivo* and *ex vivo* regulation of bacterial virulence gene expression. Curr. Opin. Microbiol. 1, 17–26.
12 MacKichan, J. K., Gaynor, E. C., Chang, C., Cawthraw, S., Newell, D. G., Miller, J. F. and Falkow, S. (2004) The *Campylobacter jejuni dccRS* two-component system is required for optimal *in vivo* colonization but is dispensable for *in vitro* growth. Mol. Microbiol. 54, 1269–1286.
13 Mandin, P., Fsihi, H., Dussurget, O., Vergassola, M., Milohanic, E., Toledo-Arana, A., Lasa, I., Johansson, J. and Cossart, P. (2005) VirR, a response regulator critical for *Listeria monocytogenes* virulence. Mol. Microbiol. 57, 1367–1380.
14 Lamy, M. C., Zouine, M., Fert, J., Vergassola, M., Couve, E., Pellegrini, E., Glaser, P., Kunst, F., Msadek, T., Trieu-Cuot, P. and Poyart, C. (2004) CovS/CovR of group B streptococcus: a two-component global regulatory system involved in virulence. Mol. Microbiol. 54, 1250–1268.
15 Dramsi, S., Bourdichon, F., Cabanes, D., Lecuit, M., Fsihi, H. and Cossart, P. (2004) FbpA, a novel multifunctional *Listeria monocytogenes* virulence factor. Mol. Microbiol. 53, 639–649.
16 Rickman, L., Scott, C., Hunt, D. M., Hutchinson, T., Menendez, M. C., Whalan, R., Hinds, J., Colston, M. J., Green, J. and Buxton, R. S. (2005) A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Mol. Microbiol. 56, 1274–1286.
17 Deziel, E., Gopalan, S., Tampakaki, A. P., Lepine, F., Padfield, K. E., Saucier, M., Xiao, G. and Rahme, L. G. (2005) The contribution of MvfR to *Pseudomonas aeruginosa* pathogenesis and quorum sensing circuitry regulation: multiple quorum sensing-regulated genes are modulated without affecting *lasRI*, *rhlRI* or the production of N-acyl-L-homoserine lactones. Mol. Microbiol. 55, 998–1014.
18 Dong, Y. H., Zhang, X. F., Xu, J. L., Tan, A. T. and Zhang, L. H. (2005) VqsM, a novel AraC-type global regulator of quorum-sensing signalling and virulence in *Pseudomonas aeruginosa*. Mol. Microbiol. 58, 552–564.
19 Stein, L. (2001) Genome annotation: from sequence to biology. Nat. Rev. Genet. 2, 493–503.
20 Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33–36.
21 Madan Babu, M. and Teichmann, S. A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res. 31, 1234–1244.
22 Doerks, T., Andrade, M. A., Lathe, W., 3rd, von Mering, C. and Bork, P. (2004) Global analysis of bacterial transcription factors to predict cellular target processes. Trends Genet. 20, 126–131.
23 Kummerfeld, S. K. and Teichmann, S. A. (2006) DBD: a transcription factor prediction database. Nucleic Acids Res. 34, D74–81.
24 Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N. et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

25 Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res. 32, D235–239.
26 Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. et al. (2004) The Pfam protein families database. Nucleic Acids Res. 32, D138–141.
27 Gadgil, H., Oak, S. A. and Jarrett, H. W. (2001) Affinity purification of DNA-binding proteins. J. Biochem. Biophys. Methods 49, 607–624.
28 Gadgil, H., Jurado, L. A. and Jarrett, H. W. (2001) DNA affinity chromatography of transcription factors. Anal. Biochem. 290, 147–178.
29 Forde, C. E. and McCutchen-Maloney, S. L. (2002) Characterization of transcription factors by mass spectrometry and the role of SELDI-MS. Mass Spectrom. Rev. 21, 419–439.
30 Nordhoff, E., Krogsdam, A. M., Jorgensen, H. F., Kallipolitis, B. H., Clark, B. F., Roepstorff, P. and Kristiansen, K. (1999) Rapid identification of DNA-binding proteins by mass spectrometry. Nat. Biotechnol. 17, 884–888.
31 Chockalingam, P. S., Gadgil, H. and Jarrett, H. W. (2002) DNA-support coupling for transcription factor purification. Comparison of aldehyde, cyanogen bromide and N-hydroxysuccinimide chemistries. J. Chromatogr. A 942, 167–175.
32 Park, S. S., Ko, B. J. and Kim, B. G. (2005) Mass spectrometric screening of transcriptional regulators using DNA affinity capture assay. Anal. Biochem. 344, 152–154.
33 Bane, T. K., LeBlanc, J. F., Lee, T. D. and Riggs, A. D. (2002) DNA affinity capture and protein profiling by SELDI-TOF mass spectrometry: effect of DNA methylation. Nucleic Acids Res. 30, e69.
34 Forde, C. E., Gonzales, A. D., Smessaert, J. M., Murphy, G. A., Shields, S. J., Fitch, J. P. and McCutchen-Maloney, S. L. (2002) A rapid method to capture and screen for transcription factors by SELDI mass spectrometry. Biochem. Biophys. Res. Commun. 290, 1328–1335.
35 Gadgil, H. and Jarrett, H. W. (2002) Oligonucleotide trapping method for purification of transcription factors. J. Chromatogr. A 966, 99–110.
36 Trubetskoy, D. O., Zavalova, L. L., Akopov, S. B. and Nikolaev, L. G. (2002) Purification of proteins specifically binding human endogenous retrovirus K long terminal repeat by affinity elution chromatography. J. Chromatogr. A 976, 95–101.
37 Yaneva, M. and Tempst, P. (2003) Affinity capture of specific DNA-binding proteins for mass spectrometric identification. Anal. Chem. 75, 6437–6448.
38 Woo, A. J., Dods, J. S., Susanto, E., Ulgiati, D. and Abraham, L. J. (2002) A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay. Mol. Cell. Proteomics 1, 472–478.
39 Hazbun, T. R. and Fields, S. (2002) A genome-wide screen for site-specific DNA-binding proteins. Mol. Cell. Proteomics 1, 538–543.
40 Martin, R. G. and Rosner, J. L. (2001) The AraC transcriptional activators. Curr. Opin. Microbiol. 4, 132–137.
41 Korner, H., Sofia, H. J. and Zumft, W. G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. FEMS Microbiol. Rev. 27, 559–592.
42 Nguyen, C. C. and Saier, M. H. Jr (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. FEBS Lett. 377, 98–102.
43 Brinkman, A. B., Ettema, T. J., de Vos, W. M. and van der Oost, J. (2003) The Lrp family of transcriptional regulators. Mol. Microbiol. 48, 287–294.
44 Schell, M. A. (1993) Molecular biology of the LysR family of transcriptional regulators. Annu. Rev. Microbiol. 47, 597–626.
45 Brown, N. L., Stoyanov, J. V., Kidd, S. P. and Hobman, J. L. (2003) The MerR family of transcriptional regulators. FEMS Microbiol. Rev. 27, 145–163.
46 Yang, Y. H., Buckley, M. J. and Speed, T. P. (2001) Analysis of cDNA microarray images. Brief. Bioinform. 2, 341–349.
47 Qin, L., Rueda, L., Ali, A. and Ngom, A. (2005) Spot detection and image segmentation in DNA microarray data. Appl. Bioinformatics 4, 1–11.
48 Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 29, 2549–2557.
49 Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, e15.
50 Quackenbush, J. (2002) Microarray data normalization and transformation. Nat. Genet. 32 Suppl, 496–501.
51 Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S. and Simon, R. (2003) Evaluation of normalization methods for microarray data. BMC Bioinformatics 4, 33.
52 Richmond, C. S., Glasner, J. D., Mau, R., Jin, H. and Blattner, F. R. (1999) Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. 27, 3821–3835.
53 Lee, P. D., Sladek, R., Greenwood, C. M. and Hudson, T. J. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res. 12, 292–297.
54 Eickhoff, B., Korn, B., Schick, M., Poustka, A. and van der Bosch, J. (1999) Normalization of array hybridization experiments in differential gene expression analysis. Nucleic Acids Res. 27, e33.
55 Benes, V. and Muckenthaler, M. (2003) Standardization of protocols in cDNA microarray analysis. Trends Biochem. Sci. 28, 244–249.
56 Smyth, G. K. and Speed, T. (2003) Normalization of cDNA microarray data. Methods 31, 265–273.
57 Rhodius, V., Van Dyk, T. K., Gross, C. and LaRossa, R. A. (2002) Impact of genomic technologies on studies of bacterial gene expression. Annu. Rev. Microbiol. 56, 599–624.
58 Novak, J. P., Sladek, R. and Hudson, T. J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. Genomics 79, 104–113.
59 Nadon, R. and Shoemaker, J. (2002) Statistical issues with microarrays: processing and analysis. Trends Genet. 18, 265–271.
60 Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000) Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. J. Biol. Chem. 275, 29672–29684.
61 Lee, M. L., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA 97, 9834–9839.
62 Dobbin, K., Shih, J. H. and Simon, R. (2003) Statistical design of reverse dye microarrays. Bioinformatics 19, 803–810.
63 Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W. and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. J. Biol. Chem. 276, 19937–19944.
64 Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA 98, 5116–5121.
65 Baldi, P. and Long, A. D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17, 509–519.

66 Kerr, M. K., Martin, M. and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. J. Comput. Biol. 7, 819–837.
67 Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S. and Tainsky, M. A. (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 19, 1348–1359.
68 Churchill, G. A. (2004) Using ANOVA to analyze microarray data. Biotechniques 37, 173–175, 177.
69 Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J. Comput. Biol. 7, 805–817.
70 Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. 8, 625–637.
71 Pan, W., Lin, J. and Le, C. T. (2003) A mixture model approach to detecting differentially expressed genes with microarray data. Funct. Integr. Genomics 3, 117–124.
72 Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18, 546–554.
73 Cui, X. and Churchill, G. A. (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4, 210.
74 Hatfield, G. W., Hung, S. P. and Baldi, P. (2003) Differential analysis of DNA microarray gene expression data. Mol. Microbiol. 47, 871–877.
75 Rhodius, V. A. and LaRossa, R. A. (2003) Uses and pitfalls of microarrays for studying transcriptional regulation. Curr. Opin. Microbiol. 6, 114–119.
76 Cases, I. and de Lorenzo, V. (2005) Promoters in the environment: transcriptional regulation in its natural context. Nat. Rev. Microbiol. 3, 105–118.
77 Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A. and Collado-Vides, J. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 34, D394–D397.
78 Sherlock, G. (2000) Analysis of large-scale gene expression data. Curr. Opin. Immunol. 12, 201–205.
79 Quackenbush, J. (2001) Computational analysis of microarray data. Nat. Rev. Genet. 2, 418–427.
80 Raychaudhuri, S., Sutphin, P. D., Chang, J. T. and Altman, R. B. (2001) Basic microarray analysis: grouping and feature reduction. Trends Biotechnol. 19, 189–193.
81 Shannon, W., Culverhouse, R. and Duncan, J. (2003) Analyzing microarray data using cluster analysis. Pharmacogenomics 4, 41–52.
82 Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868.
83 Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999) Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285.
84 Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA 96, 2907–2912.
85 Raychaudhuri, S., Stuart, J. M. and Altman, R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 455–466.
86 Ringner, M., Peterson, C. and Khan, J. (2002) Analyzing array data using supervised methods. Pharmacogenomics 3, 403–415.
87 Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA 97, 262–267.
88 Bar-Joseph, Z. (2004) Analyzing time series gene expression data. Bioinformatics 20, 2493–2503.
89 Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W. and White, K. P. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. Science 297, 2270–2275.
90 Ernst, J., Nau, G. J. and Bar-Joseph, Z. (2005) Clustering short time series gene expression data. Bioinformatics 21 Suppl. 1, i159-i168.
91 Peddada, S. D., Lobenhofer, E. K., Li, L., Afshari, C. A., Weinberg, C. R. and Umbach, D. M. (2003) Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19, 834–841.
92 De Hoon, M. J., Imoto, S. and Miyano, S. (2002) Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics 18, 1477–1485.
93 Moller-Levet, C. S., Cho, K. H. and Wolkenhauer, O. (2003) Microarray data clustering based on temporal variation: FCV with TSD preclustering. Appl. Bioinformatics 2, 35–45.
94 Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J. and Dimopoulos, G. (2005) Bayesian coclustering of Anopheles gene expression time series: study of immune defense response to multiple experimental challenges. Proc. Natl. Acad. Sci. USA 102, 16939–16944.
95 Sacchi, L., Bellazzi, R., Larizza, C., Magni, P., Curk, T., Petrovic, U. and Zupan, B. (2005) TA-clustering: cluster analysis of gene expression profiles through Temporal Abstractions. Int. J. Med. Inform. 74, 505–517.
96 Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S., Wolkenhauer, O. and Trajanoski, Z. (2005) A fully Bayesian model to cluster gene-expression profiles. Bioinformatics 21 Suppl. 2, ii130-ii136.
97 Ramoni, M. F., Sebastiani, P. and Kohane, I. S. (2002) Cluster analysis of gene expression dynamics. Proc. Natl. Acad. Sci. USA 99, 9121–9126.
98 Luan, Y. and Li, H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19, 474–482.
99 Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. and Simon, I. (2003) Continuous representations of time-series gene expression data. J. Comput. Biol. 10, 341–356.
100 Schliep, A., Schonhuth, A. and Steinhoff, C. (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19 Suppl. 1, i255–263.
101 Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K. and Jaakkola, T. S. (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc. Natl. Acad. Sci. USA 100, 10146–10151.
102 Park, T., Yi, S. G., Lee, S., Lee, S. Y., Yoo, D. H., Ahn, J. I. and Lee, Y. S. (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 19, 694–703.
103 Wichert, S., Fokianos, K. and Strimmer, K. (2004) Identifying periodically expressed transcripts in microarray time series data. Bioinformatics 20, 5–20.
104 Luan, Y. and Li, H. (2004) Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. Bioinformatics 20, 332–339.

105 Chen, J. (2005) Identification of significant periodic genes in microarray gene expression data. BMC Bioinformatics 6, 286.
106 Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005) Significance analysis of time course microarray experiments. Proc. Natl. Acad. Sci. USA 102, 12837–12842.
107 Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E. and Liu, J. S. (2004) Statistical resynchronization and Bayesian detection of periodically expressed genes. Nucleic Acids Res. 32, 447–455.
108 Ahdesmaki, M., Lahdesmaki, H., Pearson, R., Huttunen, H. and Yli-Harja, O. (2005) Robust detection of periodic time series measured from biological systems. BMC Bioinformatics 6, 117.
109 Leek, J. T., Monsen, E., Dabney, A. R. and Storey, J. D. (2006) EDGE: extraction and analysis of differential gene expression. Bioinformatics 22, 507–508.
110 Conway, T. and Schoolnik, G. K. (2003) Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Mol. Microbiol. 47, 879–889.
111 Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzel, H. (2000) Normalization strategies for cDNA microarrays. Nucleic Acids Res. 28, E47.
112 Kothapalli, R., Yoder, S. J., Mane, S. and Loughran, T. P. Jr (2002) Microarray results: how accurate are they? BMC Bioinformatics 3, 22.
113 Chuaqui, R. F., Bonner, R. F., Best, C. J., Gillespie, J. W., Flaig, M. J., Hewitt, S. M., Phillips, J. L., Krizman, D. B., Tangrea, M. A. et al. (2002) Post-analysis follow-up and validation of microarray experiments. Nat. Genet. 32 Suppl, 509–514.
114 Rajeevan, M. S., Ranamukhaarachchi, D. G., Vernon, S. D. and Unger, E. R. (2001) Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. Methods 25, 443–451.
115 Wong, M. L. and Medrano, J. F. (2005) Real-time PCR for mRNA quantitation. Biotechniques 39, 75–85.
116 Huggett, J., Dheda, K., Bustin, S. and Zumla, A. (2005) Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. 6, 279–284.
117 Lee, J. M., Zhang, S., Saha, S., Santa Anna, S., Jiang, C. and Perkins, J. (2001) RNA expression analysis using an antisense *Bacillus subtilis* genome array. J. Bacteriol. 183, 7371–7380.
118 Stintzi, A. (2003) Gene expression profile of *Campylobacter jejuni* in response to growth temperature variation. J. Bacteriol. 185, 2009–2016.
119 Smoot, L. M., Smoot, J. C., Graham, M. R., Somerville, G. A., Sturdevant, D. E., Migliaccio, C. A., Sylva, G. L. and Musser, J. M. (2001) Global differential gene expression in response to growth temperature alteration in group A *Streptococcus*. Proc. Natl. Acad. Sci. USA 98, 10416–10421.
120 Wan, X. F., Verberkmoes, N. C., McCue, L. A., Stanek, D., Connelly, H., Hauser, L. J., Wu, L., Liu, X., Yan, T., Leaphart, A., Hettich, R. L., Zhou, J. and Thompson, D. K. (2004) Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium *Shewanella oneidensis*. J. Bacteriol. 186, 8385–8400.
121 Arnone, M. I., Dmochowski, I. J. and Gache, C. (2004) Using reporter genes to study cis-regulatory elements. Methods Cell. Biol. 74, 621–652.
122 Hand, N. J. and Silhavy, T. J. (2000) A practical guide to the construction and use of *lac* fusions in *Escherichia coli*. Methods Enzymol. 326, 11–35.
123 Hamon, M. A., Stanley, N. R., Britton, R. A., Grossman, A. D. and Lazazzera, B. A. (2004) Identification of AbrB-regulated genes involved in biofilm formation by *Bacillus subtilis*. Mol. Microbiol. 52, 847–860.
124 Rowland, B., Purkayastha, A., Monserrat, C., Casart, Y., Takiff, H. and McDonough, K. A. (1999) Fluorescence-based detection of *lacZ* reporter gene expression in intact and viable bacteria including Mycobacterium species. FEMS Microbiol. Lett. 179, 317–325.
125 Slauch, J. M. and Silhavy, T. J. (1991) Genetic fusions as experimental tools. Methods Enzymol. 204, 213–248.
126 Becher, A. and Schweizer, H. P. (2000) Integration-proficient *Pseudomonas aeruginosa* vectors for isolation of single-copy chromosomal *lacZ* and *lux* gene fusions. Biotechniques 29, 948–950, 952.
127 Platt, R., Drescher, C., Park, S. K. and Phillips, G. J. (2000) Genetic system for reversible integration of DNA constructs and *lacZ* gene fusions into the *Escherichia coli* chromosome. Plasmid 43, 12–23.
128 Haldimann, A. and Wanner, B. L. (2001) Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. J. Bacteriol. 183, 6384–6393.
129 Ellermeier, C. D., Janakiraman, A. and Slauch, J. M. (2002) Construction of targeted single copy *lac* fusions using lambda Red and FLP-mediated site-specific recombination in bacteria. Gene 290, 153–161.
130 Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. and Collado-Vides, J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. Proc. Natl. Acad. Sci. USA 97, 6652–6657.
131 Yada, T., Nakao, M., Totoki, Y. and Nakai, K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. Bioinformatics 15, 987–993.
132 Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001) Prediction of operons in microbial genomes. Nucleic Acids Res. 29, 1216–1221.
133 Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. and Koonin, E. V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11, 356–372.
134 Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J. and Kasif, S. (2002) Computational identification of operons in microbial genomes. Genome Res. 12, 1221–1230.
135 Paredes, C. J., Rigoutsos, I. and Papoutsakis, E. T. (2004) Transcriptional organization of the *Clostridium acetobutylicum* genome. Nucleic Acids Res. 32, 1973–1981.
136 Wang, L., Trawick, J. D., Yamamoto, R. and Zamudio, C. (2004) Genome-wide operon prediction in *Staphylococcus aureus*. Nucleic Acids Res. 32, 3689–3702.
137 Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 Suppl. 1, S329–336.
138 Strong, M., Mallick, P., Pellegrini, M., Thompson, M. J. and Eisenberg, D. (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. Genome Biol. 4, R59.
139 Romero, P. R. and Karp, P. D. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. Bioinformatics 20, 709–717.
140 Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y. and Jiang, T. (2004) Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. Nucleic Acids Res. 32, 2147–2157.
141 Jacob, E., Sasikumar, R. and Nair, K. N. (2005) A fuzzy guided genetic algorithm for operon prediction. Bioinformatics 21, 1403–1407.
142 Edwards, M. T., Rison, S. C., Stoker, N. G. and Wernisch, L. (2005) A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. Nucleic Acids Res. 33, 3253–3262.
143 Westover, B. P., Buhler, J. D., Sonnenburg, J. L. and Gordon, J. I. (2005) Operon prediction without a training set. Bioinformatics 21, 880–888.

144 Price, M. N., Huang, K. H., Alm, E. J. and Arkin, A. P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. 33, 880–892.
145 Okuda, S., Katayama, T., Kawashima, S., Goto, S. and Kanehisa, M. (2006) ODB: a database of operons accumulating known operons across multiple genomes. Nucleic Acids Res. 34, D358–362.
146 Merrell, D. S., Thompson, L. J., Kim, C. C., Mitchell, H., Tompkins, L. S., Lee, A. and Falkow, S. (2003) Growth phase-dependent response of *Helicobacter pylori* to iron starvation. Infect. Immun. 71, 6510–6525.
147 Schuster, M., Hawkins, A. C., Harwood, C. S. and Greenberg, E. P. (2004) The *Pseudomonas aeruginosa* RpoS regulon and its relationship to quorum sensing. Mol. Microbiol. 51, 973–985.
148 Tjaden, B., Haynor, D. R., Stolyar, S., Rosenow, C. and Kolker, E. (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. Bioinformatics 18 Suppl. 1, S337–344.
149 Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F. and Craven, M. (2003) Predicting bacterial transcription units using sequence and expression data. Bioinformatics 19 Suppl. 1, i34–43.
150 Sabatti, C., Rohlin, L., Oh, M. K. and Liao, J. C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. 30, 2886–2893.
151 Yamanishi, Y., Vert, J. P., Nakaya, A. and Kanehisa, M. (2003) Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 19 Suppl. 1, i323–i330.
152 Bockhorst, J., Craven, M., Page, D., Shavlik, J. and Glasner, J. (2003) A Bayesian network approach to operon prediction. Bioinformatics 19, 1227–1235.
153 Steinhauser, D., Junker, B. H., Luedemann, A., Selbig, J. and Kopka, J. (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. Bioinformatics 20, 1928–1939.
154 Gupta, A. (1999) RT-PCR: characterization of long multigene operons and multiple transcript gene clusters in bacteria. Biotechniques 27, 966–970, 972.
155 Guacucano, M., Levican, G., Holmes, D. S. and Jedlicki, E. (2000) An RT-PCR artifact in the characterization of bacterial operons. Electro. J. Biotechnol. 3, 1–4.
156 Orlando, V. (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-cross-linked-chromatin immunoprecipitation. Trends Biochem. Sci. 25, 99–104.
157 Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290, 2306–2309.
158 Spencer, V. A., Sun, J. M., Li, L. and Davie, J. R. (2003) Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. Methods 31, 67–75.
159 Steward, N. and Sano, H. (2004) Measuring changes in chromatin using micrococcal nuclease. Methods Mol. Biol. 287, 65–75.
160 Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. and Brown, P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409, 533–538.
161 Laub, M. T., Chen, S. L., Shapiro, L. and McAdams, H. H. (2002) Genes directly controlled by CtrA, a master regulator of the Caulobacter cell cycle. Proc. Natl. Acad. Sci. USA 99, 4632–4637.
162 Bohlander, S. K., Espinosa, R., 3rd, Le Beau, M. M., Rowley, J. D. and Diaz, M. O. (1992) A method for the rapid sequence-independent amplification of microdissected chromosomal material. Genomics 13, 1322–1324.
163 Mueller, P. R. and Wold, B. (1989) *In vivo* footprinting of a muscle specific enhancer by ligation mediated PCR. Science 246, 780–786.
164 Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. and Busby, S. J. (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the E. coli chromosome. Proc. Natl. Acad. Sci. USA 102, 17693–17698.
165 Xiang, C. C., Kozhich, O. A., Chen, M., Inman, J. M., Phan, Q. N., Chen, Y. and Brownstein, M. J. (2002) Amine-modified random primers to label probes for DNA microarrays. Nat. Biotechnol. 20, 738–742.
166 Sikder, D. and Kodadek, T. (2005) Genomic studies of transcription factor-DNA interactions. Curr. Opin. Chem. Biol 9, 38–45.
167 Hanlon, S. E. and Lieb, J. D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. Curr. Opin. Genet. Dev. 14, 697–705.
168 Buck, M. J. and Lieb, J. D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83, 349–360.
169 Molle, V., Nakaura, Y., Shivers, R. P., Yamaguchi, H., Losick, R., Fujita, Y. and Sonenshein, A. L. (2003) Additional targets of the *Bacillus subtilis* global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis. J. Bacteriol. 185, 1911–1922.
170 Molle, V., Fujita, M., Jensen, S. T., Eichenberger, P., Gonzalez-Pastor, J. E., Liu, J. S. and Losick, R. (2003) The Spo0A regulon of *Bacillus subtilis*. Mol. Microbiol. 50, 1683–1701.
171 Grainger, D. C., Overton, T. W., Reppas, N., Wade, J. T., Tamai, E., Hobman, J. L., Constantinidou, C., Struhl, K., Church, G. and Busby, S. J. (2004) Genomic studies with *Escherichia coli* MelR protein: applications of chromatin immunoprecipitation and microarrays. J. Bacteriol. 186, 6938–6943.
172 Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T., Ferguson, C., Haga, K., Sato, T., Liu, J. S. and Losick, R. (2004) The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. PLoS Biol. 2, e328.
173 Herring, C. D., Raffaelle, M., Allen, T. E., Kanin, E. I., Landick, R., Ansari, A. Z. and Palsson, B. O. (2005) Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. J. Bacteriol. 187, 6166–6174.
174 Buck, M. J., Nobel, A. B. and Lieb, J. D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol. 6, R97.
175 Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M., Rinn, J. L., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. Proc. Natl. Acad. Sci. USA 100, 12247–12252.
176 Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. (2004) CREB binds to multiple loci on human chromosome 22. Mol. Cell. Biol. 24, 3804–3814.
177 Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5, 276–287.
178 Li, H., Rhodius, V., Gross, C. and Siggia, E. D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. Proc. Natl. Acad. Sci. USA 99, 11772–11777.
179 van Helden, J., Rios, A. F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 28, 1808–1818.

180 Vanet, A., Marsan, L., Labigne, A. and Sagot, M. F. (2000) Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals. J. Mol. Biol. 297, 335–353.
181 Bussemaker, H. J., Li, H. and Siggia, E. D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc. Natl. Acad. Sci. USA 97, 10096–10100.
182 Stormo, G. D. (2000) DNA binding sites: representation and discovery. Bioinformatics 16, 16–23.
183 van Helden, J. (2003) Regulatory sequence analysis tools. Nucleic Acids Res. 31, 3593–3596.
184 Zhou, D., Qin, L., Han, Y., Qiu, J., Chen, Z., Li, B., Song, Y., Wang, J., Guo, Z., Zhai, J. et al. (2006) Global analysis of iron assimilation and Fur regulation in *Yersinia pestis*. FEMS Microbiol. Lett. 258, 9–17.
185 Mao, L., Mackenzie, C., Roh, J. H., Eraso, J. M., Kaplan, S. and Resat, H. (2005) Combining microarray and genomic data to predict DNA binding motifs. Microbiology 151, 3197–3213.
186 Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y. and De Moor, B. (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res. 33, W393–396.
187 Conlon, E. M., Liu, X. S., Lieb, J. D. and Liu, J. S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. Proc. Natl. Acad. Sci. USA 100, 3339–3344.
188 Liu, X. S., Brutlag, D. L. and Liu, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. 20, 835–839.
189 Hong, P., Liu, X. S., Zhou, Q., Lu, X., Liu, J. S. and Wong, W. H. (2005) A boosting approach for motif modeling using ChIP-chip data. Bioinformatics 21, 2636–2643.
190 Smith, A. D., Sumazin, P., Das, D. and Zhang, M. Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. Bioinformatics 21 Suppl. 1, i403-i412.
191 Li, W., Meyer, C. A. and Liu, X. S. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 21 Suppl. 1, i274-i282.
192 Hu, J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. 33, 4899–4913.
193 Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. 23, 137–144.
194 Wingender, E., Dietze, P., Karas, H. and Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24, 238–241.
195 Guia, M. H., Perez, A. G., Angarica, V. E., Vasconcelos, A. T. and Collado-Vides, J. (2005) Complementing computationally predicted regulatory sites in Tractor_DB using a pattern matching approach. In Silico Biol. 5, 209–219.
196 Wasserman, W. W. and Fickett, J. W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. 278, 167–181.
197 Frech, K., Herrmann, G. and Werner, T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. Nucleic Acids Res. 21, 1655–1664.
198 Hoglund, A. and Kohlbacher, O. (2004) From sequence to structure and back again: approaches for predicting protein-DNA binding. Proteome Sci. 2, 3.
199 Zheng, M., Wang, X., Doan, B., Lewis, K. A., Schneider, T. D. and Storz, G. (2001) Computation-directed identification of OxyR DNA binding sites in *Escherichia coli*. J. Bacteriol. 183, 4571–4579.
200 Groisman, E. A. (2001) The pleiotropic two-component regulatory system PhoP-PhoQ. J Bacteriol 183, 1835–1842.
201 Lejona, S., Aguirre, A., Cabeza, M. L., Garcia Vescovi, E. and Soncini, F. C. (2003) Molecular characterization of the Mg^{2+}-responsive PhoP-PhoQ regulon in *Salmonella enterica*. J. Bacteriol. 185, 6287–6294.
202 Lane, D., Prentki, P. and Chandler, M. (1992) Use of gel retardation to analyze protein-nucleic acid interactions. Microbiol. Rev. 56, 509–528.
203 Carey, J. (1991) Gel retardation. Methods Enzymol. 208, 103–117.
204 Molloy, P. L. (2000) Electrophoretic mobility shift assays. Methods Mol. Biol. 130, 235–246.
205 Yoshimura, H., Yanagisawa, S., Kanehisa, M. and Ohmori, M. (2002) Screening for the target gene of cyanobacterial cAMP receptor protein SYCRP1. Mol. Microbiol. 43, 843–853.
206 Meyer-ter-Vehn, T., Covacci, A., Kist, M. and Pahl, H. L. (2000) *Helicobacter pylori* activates mitogen-activated protein kinase cascades and induces expression of the proto-oncogenes c-fos and c-jun. J. Biol. Chem. 275, 16064–16072.
207 Galas, D. J. and Schmitz, A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 5, 3157–3170.
208 Baichoo, N., Wang, T., Ye, R. and Helmann, J. D. (2002) Global analysis of the Bacillus subtilis Fur regulon and the iron starvation stimulon. Mol. Microbiol. 45, 1613–1629.
209 Holden, C. (2002) Cell biology. Alliance launched to model *E. coli*. Science 297, 1459–1460.
210 Engels, S., Schweitzer, J. E., Ludwig, C., Bott, M. and Schaffer, S. (2004) *clpC* and *clpP1P2* gene expression in *Corynebacterium glutamicum* is controlled by a regulatory network involving the transcriptional regulators ClgR and HspR as well as the ECF sigma factor sigmaH. Mol. Microbiol. 52, 285–302.
211 Minagawa, S., Ogasawara, H., Kato, A., Yamamoto, K., Eguchi, Y., Oshima, T., Mori, H., Ishihama, A. and Utsumi, R. (2003) Identification and molecular characterization of the Mg^{2+} stimulon of *Escherichia coli*. J. Bacteriol. 185, 3696–3702.
212 Park, S. H. and Raines, R. T. (2004) Fluorescence gel retardation assay to detect protein-protein interactions. Methods Mol. Biol. 261, 155–160.
213 Feriotto, G., Mischiati, C., Bianchi, N., Passadore, M. and Gambari, R. (1995) Binding of distamycin and chromomycin to human immunodeficiency type 1 virus DNA: a non-radioactive automated footprinting study. Eur. J. Pharmacol. 290, 85–93.
214 Yamada, M., Izu, H., Nitta, T., Kurihara, K. and Sakurai, T. (1998) High-temperature, nonradioactive primer extension assay for determination of a transcription-initiation site. Biotechniques 25, 72–74, 76, 78.
215 Machida, M., Kamio, H. and Sorensen, D. (1997) Long-range and highly sensitive DNase I footprinting by an automated infrared DNA sequencer. Biotechniques 23, 300–303.
216 Wilson, D. O., Johnson, P. and McCord, B. R. (2001) Nonradiochemical DNase I footprinting by capillary electrophoresis. Electrophoresis 22, 1979–1986.
217 Galperin, M. Y. (2006) The Molecular Biology Database Collection: 2006 update. Nucleic Acids Res. 34, D3–5.
218 Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283–291.
219 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: simple building blocks of complex networks. Science 298, 824–827.
220 Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31, 64–68.

221 Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.

222 Wang, E. and Purisima, E. (2005) Network motifs are enriched with transcription factors whose transcripts have short half-lives. Trends Genet. 21, 492–495.

223 Prill, R. J., Iglesias, P. A. and Levchenko, A. (2005) Dynamic properties of network motifs contribute to biological network organization. PLoS Biol. 3, e343.

224 Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M. and Karp, P. D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res. 33, D334–337.

225 Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Penaloza-Spinola, M. I., Martinez-Antonio, A., Karp, P. D. and Collado-Vides, J. (2006) The comprehensive updated regulatory network of *Escherichia coli* K-12. BMC Bioinformatics 7, 5.

226 Ma, H. W., Buer, J. and Zeng, A. P. (2004) Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics 5, 199.

227 Ma, H. W., Kumar, B., Ditges, U., Gunzer, F., Buer, J. and Zeng, A. P. (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 32, 6643–6649.

228 Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100, 11980–11985.

229 Dobrin, R., Beg, Q. K., Barabasi, A. L. and Oltvai, Z. N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. BMC Bioinformatics 5, 10.

230 Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B. L., Mori, H. and Mizuno, T. (2002) Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12. Mol. Microbiol. 46, 281–291.

231 Gutierrez-Rios, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R. and Collado-Vides, J. (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. Genome Res. 13, 2435–2443.

232 Herrgard, M. J., Covert, M. W. and Palsson, B. O. (2003) Reconciling gene expression data with known genome-scale regulatory network structures. Genome Res. 13, 2423–2434.

233 Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19, 2271–2282.

234 D'Haeseleer, P., Liang, S. and Somogyi, R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726.

235 Liang, S., Fuhrman, S. and Somogyi, R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac. Symp. Biocomput., 18–29.

236 Chen, T., He, H. L. and Church, G. M. (1999) Modeling gene expression with differential equations. Pac. Symp. Biocomput., 29–40.

237 Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601–620.

238 van Someren, E. P., Wessels, L. F., Backer, E. and Reinders, M. J. (2002) Genetic network modeling. Pharmacogenomics 3, 507–525.

239 de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 9, 67–103.

240 Bolouri, H. and Davidson, E. H. (2002) Modeling transcriptional regulatory networks. Bioessays 24, 1118–1129.

241 Wang, W., Cherry, J. M., Botstein, D. and Li, H. (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 99, 16893–16898.

242 Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003) Computational discovery of gene modules and regulatory networks. Nat. Biotechnol. 21, 1337–1342.

243 Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34, 166–176.

244 Schuller, C., Mamnun, Y. M., Mollapour, M., Krapf, G., Schuster, M., Bauer, B. E., Piper, P. W. and Kuchler, K. (2004) Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in *Saccharomyces cerevisiae*. Mol. Biol. Cell. 15, 706–720.

245 Blais, A. and Dynlacht, B. D. (2005) Constructing transcriptional regulatory networks. Genes Dev. 19, 1499–1511.

246 Murakami, K. S. and Darst, S. A. (2003) Bacterial RNA polymerases: the wholo story. Curr. Opin. Struct. Biol. 13, 31–39.

247 Gourse, R. L., Ross, W. and Gaal, T. (2000) UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. Mol. Microbiol. 37, 687–695.

248 Lawson, C. L., Swigon, D., Murakami, K. S., Darst, S. A., Berman, H. M. and Ebright, R. H. (2004) Catabolite activator protein: DNA binding and transcription activation. Curr. Opin. Struct. Biol. 14, 10–20.

249 Wonderling, L. D. and Stauffer, G. V. (1999) The cyclic AMP receptor protein is dependent on GcvA for regulation of the *gcv* operon. J. Bacteriol. 181, 1912–1919.

250 Staden, R. (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12, 505–519.

251 Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100.

252 Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

253 Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003) Global functional profiling of gene expression. Genomics 81, 98–104.

254 Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M. et al. (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378.

255 Bailey, T. L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the International Conference on Intelligent Systems for Molecular Biology, ISMB 2, 28–36.

256 Hertz, G. Z. and Stormo, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563–577.

257 Thompson, W., Rouchka, E. C. and Lawrence, C. E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res. 31, 3580–3585.

258 Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. 16, 939–945.

259 Hughes, J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J. Mol. Biol. 296, 1205–1214.

260 Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2001) A higher-order back-

ground model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17, 1113–1122.
261 Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J. Comput. Biol. 9, 447–464.
262 Liu, X., Brutlag, D. L. and Liu, J. S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput., 127–138.
263 Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32, W199–W203.
264 Zhou, Q. and Wong, W. H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc. Natl. Acad. Sci. USA 101, 12114–12119.
265 Olson, S. A. (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief. Bioinform. 3, 87–91.
266 Bailey, T. L. and Gribskov, M. (1998) Methods and statistics for combining motif match scores. J. Comput. Biol. 5, 211–221.
267 Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23, 4878–4884.
268 Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. and Wingender, E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31, 3576–3579.
269 Chekmenev, D. S., Haid, C. and Kel, A. E. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. Nucleic Acids Res. 33, W432–W437.
270 Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. and Pesole, G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Res. 31, 3608–3612.
271 Chaudhuri, R. R., Khan, A. M. and Pallen, M. J. (2004) *coli*BASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. Nucleic Acids Res. 32, D296–D299.
272 Johansson, O., Alkema, W., Wasserman, W. W. and Lagergren, J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics 19 Suppl. 1, i169–i176.
273 Fu, Y., Frith, M. C., Haverty, P. M. and Weng, Z. (2004) MotifViz: an analysis and visualization tool for motif discovery. Nucleic Acids Res. 32, W420–W423.
274 Workman, C. T., Yin, Y., Corcoran, D. L., Ideker, T., Stormo, G. D. and Benos, P. V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. 33, W389–W392.
275 Coessens, B., Thijs, G., Aerts, S., Marchal, K., De Smet, F., Engelen, K., Glenisson, P., Moreau, Y., Mathys, J. and De Moor, B. (2003) INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res. 31, 3468–3470.
276 Okubo, K., Sugawara, H., Gojobori, T. and Tateno, Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. Nucleic Acids Res. 34, D6–D9.
277 Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A. et al. (2006) EMBL Nucleotide Sequence Database: developments in 2005. Nucleic Acids Res. 34, D10–D15.
278 Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2006) GenBank. Nucleic Acids Res. 34, D16–D20.
279 Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G. et al. (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31, 68–71.
280 Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C. et al. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. 31, 94–96.
281 Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L. and Tobes, R. (2004) BacTregulators: a database of transcriptional regulators in bacteria and archaea. Bioinformatics 20, 2787–2791.
282 Munch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E. and Jahn, D. (2003) PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res. 31, 266–269.
283 Gonzalez, A. D., Espinosa, V., Vasconcelos, A. T., Perez-Rueda, E. and Collado-Vides, J. (2005) TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. Nucleic Acids Res. 33, D98–D102.
284 Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E. et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 33, D418–D424.
285 Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Res. 32, D75–D77.
286 Jacques, P. E., Gervais, A. L., Cantin, M., Lucier, J. F., Dallaire, G., Drouin, G., Gaudreau, L., Goulet, J. and Brzezinski, R. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosi*s. Bioinformatics 21, 2563–2565.
287 Robison, K., McGuire, A. M. and Church, G. M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284, 241–254.
288 Hershberg, R., Bejerano, G., Santos-Zavaleta, A. and Margalit, H. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. Nucleic Acids Res. 29, 277.