# 3

# Guide to Genome-Wide Bacterial Transcription Factor Binding Site Prediction Using OmpR as Model

Phu Vuong and Rajeev Misra
*Arizona State University*
*USA*

## 1. Introduction

Gene expression regulation in a cell plays a crucial role in the cellular response to environmental cues and other important biological processes (Bauer et al., 2010). A major mechanism of gene expression regulation is the binding of transcription factor (TF) protein to a specific DNA sequence in the regulatory region of a gene, thereby activating or inhibiting its transcription (Zhou & Liu, 2004). A TF often regulates multiple genes whose binding sites have similar but not identical sequences (Zhang et al., 2009). There is, however, a short, recurring pattern among the promoter sequences called a motif, and it is this motif that a TF recognizes and interacts with (D'haeseleer, 2006b). It is important to identify the set of genes a TF modulates, called its regulon, as this will advance our understanding of the regulatory network of an organism (D'haeseleer, 2006b; Tan et al., 2005). One way to identify the regulon is to determine a TF's motif and subsequently use the motif to search for other candidate genes regulated by the TF.

Traditionally, TF binding sites (TFBSs) are determined by various experimental approaches. Mutagenesis, DNase footprinting, gel-shift, and reporter construct assays are common methods for identifying the binding sites upstream of individual genes, but the throughput of these techniques is low (D'haeseleer, 2006b; Ladunga, 2010). In recent years, chromatin immunoprecipitation (ChIP) and systematic evolution of ligands by exponential enrichment (SELEX) are available to study protein-DNA interactions in a high throughput manner. Chromatin-immunoprecipitation of DNA cross-linked to a TF can be hybridized to a microarray (ChIP-chip) or sequenced (ChIP-seq) to obtain the TF's cognate binding sites on the whole genomic scale (Homann and Johnson, 2010; Ladunga, 2010; Stormo, 2010). SELEX is an *in vitro* technique that measures the binding affinities of TFs for synthetic, randomly generated oligonucleotides, usually 10-30 bp long (D'haeseleer, 2006b; Ladunga, 2010; Stormo, 2010). Sequences that strongly bind to a TF in question will be selectively amplified for later identification (Schug, 2008).

The major drawback of experimental approaches to determine TF recognition motifs is the time required and the relative high cost (Zhou & Liu, 2004). Moreover, some methods have specific requirements. For example, ChIP requires antibodies and certain growth conditions under which the transcription regulator is active (Tan et al., 2005). Even if a biologist can satisfy the requirements, the resolution of the regions containing the binding sites can span

from 30-50 bp (for SELEX) to a few hundreds bp (for ChIP-chip), making the extraction of the consensus motif from the collected sequences not an easy task (Stormo, 2010).
Thankfully, in recent times, a new, *in silico* approach to identify TF binding sites became available. Many bioinformatics programs—the number ranges from 120 (Wei & Yu, 2007) to over 200 (Ladunga, 2010)—have been created to help biologists predict DNA binding motifs from the enormous amounts of sequence and gene expression data generated from advances in high-throughput genomic sequencing and gene expression analysis techniques.

### 1.1 Pattern matching and pattern discovery

There are two types of motif searches and the type dictates which programs one uses. In the first type, known as unsupervised motif finding or *de novo*, *ab initio*, or pattern discovery, a researcher wants to know the consensus pattern in a set of orthologous genes, genes in a common pathway, or transcriptionally co-regulated genes or operons from an experiment (Mrazek, 2009). The genes presumably share some binding sequence for a common TF and the task is to discover the conserved, statistically over-represented motif in the regulatory regions (Mrazek, 2009). In the second type, known as supervised motif finding or pattern matching, the DNA binding motif for a TF has been determined—either predicted *de novo* or experimentally identified—and the goal is to find which other genes in the genome have a similar motif in their promoter (Mrazek, 2009).
Because pattern discovery and pattern matching are fundamentally different tasks, there are two classes of motif prediction programs, each implementing different algorithms to solve their respective problem. For *de novo* motif discovery programs, the goal is to iteratively find a set of 12-20-bp sequence motifs that are most significantly similar to each other (Mrazek, 2009), usually with an enumeration, expectation maximization, or Gibbs sampling algorithm. Representative programs of this class include AlignACE, MEME, BioProspector, MDScan and MotifSampler (Hu et al., 2005). This is the extent of our coverage on pattern discovery in this chapter. For more information, see Ladunga, 2010; MacIsaac & Fraenkel, 2006; Mrazek, 2009; Stormo, 2000; and Wei & Yu, 2007. For more details on the algorithms, see Das & Dai, 2007; D'haeseleer, 2006a; Pavesi et al., 2004; and Stormo, 2010.

### 1.2 What's covered

The rest of this chapter discusses pattern matching with a focus on prokaryotes. Eukaryotes are not covered because transcription regulation is substantially different between these two groups (Quest et al., 2008). Promoters of prokaryotes are typically less than 500 bp and are more likely to be palindromic, whereas those in eukaryotes can extend tens of thousands of nucleotides (Thompson et al., 2007). Another difference is that prokaryotic TFBSs are a few hundred bp upstream of translational start site and can overlap or appear in tandem, whereas in eukaryotes, they can be kilobases away (Bulyk, 2003; Yanover et al., 2009). Lastly, in prokaryotes, gene regulation occurs mainly at the transcriptional level (Yanover et al., 2009). In eukaryotes, multiple TFs coordinately bind to relatively short binding sites in the promoter of a single gene to regulate its expression (Thompson et al., 2007; Yoshida et al., 2006; Zaslavsky & Singh, 2006). Also in eukaryotes, the genome is bigger with more non-coding sequences and the regulatory elements can be located upstream of the gene, within it, or even downstream of it (Bulyk, 2003). With eukaryote gene regulation being more complex, motif finding programs work significantly better on lower organisms than on higher organisms (Das & Dai, 2007).

The chapter is intended to be a pragmatic guide for microbiologists. As such, it does not cover algorithms in details and technical mathematical formulas. Instead, it presents a high-level conceptual overview of the key concepts researchers need to know in order to effectively use the available bioinformatics tools to locate TF binding sites in sequenced prokaryote genomes. Online databases of prokaryote gene expression regulation information are introduced next, followed by pattern matching programs designed for or tested on prokaryotes. Finally, the chapter concludes by offering practical strategies and tips to improve the specificity and sensitivity of the results.
Along the way, the global transcriptional regulator OmpR in *Escherichia coli* will be used in examples throughout the chapter. OmpR is a cytosolic response regulator, and together with the membrane-bound histidine sensor kinase EnvZ, constitute a prototypical two-component signal transduction system in bacteria. Our lab is currently using bioinformatics to identify novel target genes of OmpR in the *E. coli* genome.

## 2. Motif representation

As mentioned before, a TF binds to different DNA sequence variations to modulate the expression of their target genes. This degeneracy of the binding sequences allows different levels of gene regulation to be achieved (D'haeseleer, 2006b). For instance, OmpR's DNA binding properties fluctuate with the extent of covalent modifications, leading to changes in the DNA binding affinity and/or its DNA binding "signature", and thus broadening its motif definition. In *E. coli* one of the genes regulated by OmpR, *ompF*, illustrates the transcriptional regulator's broad recognition signature. *In vivo* and *in vitro* experiments have shown that two OmpR molecules bind to each of the four sites in the promoter of the *ompF* gene in a tandem manner (Harlocker et al., 1995; Yoshida et al., 2006):

| | |
|---|---|
| F1: | TTTACTTTTGGTTACATATT |
| F2: | TTTTCTTTTTGAAACCAAAT |
| F3: | TTATCTTTGTAGCACTTTCA |
| F4: | GTTACGGAATATTACATTGC |

Many pattern matching programs take a set of TFBS sequences, such as the OmpR binding sequences above, as input and internally convert it to a matrix representation for genome scanning. A few programs, such as MAST, require the matrix directly, which can be constructed using one of the matrix utility programs discussed later.
Conceptually, a motif matrix is a table of 4 rows by *n* columns, where *n* is the length of the TFBS sequences, that tabulates the frequency information of the nucleotides at each position. The four rows correspond to the four nucleotides A, T, G, C. Each column in the table holds the occurrence frequency of each base at that motif position. Bases that occur more frequently at a position/column have a higher number. See Fig. 1(a) for the matrix representation of the four F1-F4 OmpR binding sequences shown above.
The matrix in Fig. 1(a) is called a position frequency matrix. In actual practice, bioinformatics programs add values like pseudocounts (to avoid zero, which is undefined for some mathematical functions used in the algorithm) and background model probabilities (to account for genome differences like GC content) to each frequency number

(Mrazek, 2009). It is this more sophisticated matrix, called a position-specific score matrix (PSSM) or position weight matrix (PWM), that is actually used by pattern matching programs during genome scanning. See Fig. 1(b).

Matrices are generally used to represent more degenerate (that is, less conserved) TFBS sequences (Mrazek, 2009). When the consensus pattern is more conserved, one may model the motif using the International Union of Pure and Applied Chemistry (IUPAC) codes (D'haeseleer, 2006a). The IUPAC system defines 11 new single-letter codes that represent more than one nucleotide (see Table 1). For example, the *ompF* binding sites could be neatly represented using the IUPAC alphabet as KTWWCKKWDKRDHACHWWNH [see Fig. 1(c)]. The *K* is a "wild card" code for guanine or thymine; *W*, adenine or thymine; and so on.

| IUPAC code | Matches Nucleotide(s) |
|---|---|
| A | A |
| C | C |
| G | G |
| T | T |
| R | A or G |
| Y | C or T |
| W | A or T |
| S | G or C |
| M | A or C |
| K | G or T |
| H | A, C or T |
| B | C, G or T |
| V | A, C or G |
| D | A, G or T |
| N | A, C, G or T |
| . or - | (gap) |

Table 1. IUPAC codes for describing more conserved transcription factor binding consensus sequences (Pavesi et al., 2004).

Another way to represent more conserved motifs is via regular expression, or regex, a complex but highly flexible language for describing text patterns in the computer field (Mrazek, 2009). One simple regular expression that describes the four OmpR-binding sites upstream of *ompF* is: [TG]T[AT][AT]C[TG][TG][AT][ATG][TG][GA][ATG][ATC]AC [ATC][AT][AT][ATGC][ATC]

Each pair of brackets specifies one nucleotide and the bases in the brackets specify the allowable nucleotides. There are slightly different flavors of the regex language that implement slightly different features, so be sure to check the documentation accompanying a pattern matching program to find out the features supported.

Note that both IUPAC codes and regular expression allow multiple bases to be specified at a nucleotide position, but all the valid bases are assumed to occur with the same frequency. Because the set of DNA sequences recognized by a TF is often degenerate and nucleotide frequency information is helpful in pattern matching, matrices are more often used and are supported by many programs.

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | T | T | T | A | C | T | T | T | T | G | G | T | T | A | C | A | T | A | T | T |
| F2 | T | T | T | T | C | T | T | T | T | T | G | A | A | A | C | C | A | A | A | T |
| F3 | T | T | A | T | C | T | T | T | G | T | A | G | C | A | C | T | T | T | C | A |
| F4 | G | T | T | A | C | G | G | A | A | T | A | T | T | A | C | A | T | T | G | C |
| (a) A | 0.00 | 0.00 | 0.25 | 0.50 | 0.00 | 0.00 | 0.00 | 0.25 | 0.25 | 0.00 | 0.50 | 0.25 | 0.25 | 1.00 | 0.00 | 0.50 | 0.25 | 0.50 | 0.25 | 0.25 |
| T | 0.75 | 1.00 | 0.75 | 0.50 | 0.00 | 0.75 | 0.75 | 0.75 | 0.50 | 0.75 | 0.00 | 0.50 | 0.50 | 0.00 | 0.00 | 0.25 | 0.75 | 0.50 | 0.25 | 0.50 |
| C | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.00 | 1.00 | 0.25 | 0.00 | 0.00 | 0.25 | 0.25 |
| G | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.25 | 0.00 | 0.25 | 0.25 | 0.50 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.00 |
| (b) A | -8.65 | -8.65 | -0.54 | 0.46 | -8.65 | -8.65 | -8.65 | -0.54 | -0.54 | -8.65 | 0.46 | -0.54 | -0.54 | 1.46 | -8.65 | 0.46 | -0.54 | 0.46 | -0.54 | -0.54 |
| T | 1.04 | 1.46 | 1.04 | 0.46 | -8.65 | 1.04 | 1.04 | 1.04 | 0.46 | 1.04 | -8.65 | 0.46 | 0.46 | -8.65 | -8.65 | -0.54 | 1.04 | 0.46 | -0.54 | 0.46 |
| C | -8.65 | -8.65 | -8.65 | -8.65 | 2.87 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | -8.65 | 2.87 | 0.87 | -8.65 | -8.65 | 0.87 | 0.87 |
| G | 0.87 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | 0.87 | -8.65 | 0.87 | 0.87 | 1.87 | 0.87 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | -8.65 |
| (c) IUPAC | K | T | W | W | C | K | K | W | D | K | R | D | H | A | C | H | W | W | N | H |

Fig. 1. The first four rows, labeled F1-F4, contain the four sites upstream of the *ompF* gene where OmpR binds in *Escherichia coli*. (a) The position frequency matrix representation of the same *ompF* sequences. Each column contains the frequencies of occurrence of the nucleotides in each corresponding F1-F4 sequence position. (b) The position weight matrix representation of the *ompF* F1-F4 sequences. The weight matrix is derived from the frequency matrix in (a). The values are calculated by taking the log of the frequency values divided by background model values. It is this position weight matrix that is actually used by pattern matching programs during execution. (c) The consensus motif of the four F1-F4 sequences in IUPAC codes.

## 3. How pattern matching programs work

Whether a motif is given as a regular expression, an IUPAC consensus sequence, or a matrix, a pattern matching program looks for the motif by scanning a genome on both the sense and antisense strands from the 5′ to 3′ end (MacIsaac & Fraenkel, 2006). See Fig. 2. Typically, the default is to check only the intergenic regions; coding regions are skipped over. A window with a width equal to the length of the motif slides over the genome one base at a time. At each iteration, the sequence in the window is checked against the given motif for a match.

For regular expression or IUPAC motif, each nucleotide in the window is checked to see if that nucleotide is allowed at that position. If the number of mismatches is at or below a certain limit, the sequence is considered a match and returned. If the motif is given as a matrix, the sequence is scored against the matrix. The score for that sequence is calculated by summing the weight score at each position. If the score is at or above a certain threshold, that sequence is considered similar to the motif and a match is found. The score measures how closely the candidate sequence matches the motif modeled by the position weight matrix and how likely the candidate happens to be a random genomic background sequence.

## 4. Motif databases and utilities

To use motif matching programs to discover candidate genes modulated by a TF, the TF's motif is required. One can look in the literature to compile a list of the reported binding site

sequences, or better yet, one can search online databases of sequenced genomes and gene regulation information, usually curated from primary journals. There are general databases covering the prokaryotes and specialized ones for particular bacterial strains (see Table 2). For instance, PRODORIC contains close to 3,000 TFBSs and over 2,000 genes for multiple bacteria species (Grote et al., 2009). Another resource containing information on transcription factors and their target genes, but for *Escherichia coli* K-12 only, is RegulonDB (Gama-Castro et al., 2011).

... CTTTTTGAAAATACGCAACGGCCATTTTTTGCACTTAGATACAGATTTTCTGCGCTGTATTGCATTGATTTGAT...

| a) | T | T | T | G | C | A | C | T | T | A | G | A | T | A | C | A | G | A | T | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | K | T | W | W | C | K | K | W | D | K | R | D | H | A | C | H | W | W | N | H |
| | Y | Y | Y | N | Y | N | N | Y | Y | N | Y | Y | Y | Y | Y | Y | N | Y | Y | Y |

| b) | T | T | T | G | C | A | C | T | T | A | G | A | T | A | C | A | G | A | T | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | -8.65 | -8.65 | -0.54 | 0.46 | -8.65 | -8.65 | -8.65 | -0.54 | -0.54 | -8.65 | 0.46 | -0.54 | -0.54 | 1.46 | -8.65 | 0.46 | -0.54 | 0.46 | -0.54 | -0.54 |
| T | 1.04 | 1.46 | 1.04 | 0.46 | -8.65 | 1.04 | 1.04 | 1.04 | 0.46 | 1.04 | -8.65 | 0.46 | 0.46 | -8.65 | -8.65 | -0.54 | 1.04 | 0.46 | -0.54 | 0.46 |
| C | -8.65 | -8.65 | -8.65 | -8.65 | 2.87 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | -8.65 | 2.87 | 0.87 | -8.65 | -8.65 | 0.87 | 0.87 |
| G | 0.87 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | 0.87 | -8.65 | 0.87 | 0.87 | 1.87 | 0.87 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | -8.65 | 0.87 | -8.65 |

Score: 1.04+1.46+1.04-8.65+2.87-8.65-8.65+1.04+0.46-8.65+1.87-0.54+0.46+1.46+2.87+0.46-8.65+0.46-0.54+0.46 = -28.38

Fig. 2. Overview of how pattern matching programs work given a motif represented using (a) IUPAC codes or (b) a position weight matrix. These programs slide an *n*-bp window (shaded and underlined), where *n* is the length of the motif (*n* = 20 in this example), over every single base in the genome. For each iteration, the sequence inside the window is (a) compared against the allowable nucleotides specified by the IUPAC motif, and if the number of mismatches is at or below a certain limit, the sequence is considered a match. For the matrix in (b), the score of an individual base in each column is looked up and summed, and if the total is at or over a certain threshold, the sequence is considered a match. The IUPAC consensus motif and the matrix are the same as those in Fig. 1(c) and (b). In (a), *Y* = a match, *N* = mismatch.

| Database | Organism | Web Address | Reference |
|---|---|---|---|
| DBTBS | *Bacillus subtilis* | http://dbtbs.hgc.jp | Sierro et al., 2008 |
| DPInteract | *Escherichia coli* | http://arep.med.harvard.edu/dpinteract | Robison et al., 1998 |
| RegulonDB | *Escherichia coli* | http://regulondb.ccg.unam.mx | Gama-Castro et al., 2011 |
| CoryneRegNet | *Corynebacterium* | http://www.CoryneRegNet.de | Baumbach, 2007 |
| PRODORIC | *Prokaryotes* | http://www.prodoric.de | Grote et al., 2009 |
| RegPrecise | *Prokaryotes* | http://regprecise.lbl.gov/RegPrecise | Novichkov et al., 2010a |
| RegTransBase | *Prokaryotes* | http://regtransbase.lbl.gov | Kazakov et al., 2007 |
| KEGG | *Prokaryotes* | http://www.genome.ad.jp/kegg | Kanehisa et al., 2004 |

Table 2. Select databases of curated and annotated transcription factor binding sites and other gene expression regulation information in prokaryotes. Note that the KEGG database also covers eukaryotes.

Once a set of binding site sequences has been gathered, online tools are available to analyze and display the motif in those sequences (Table 3). D-MATRIX (Sen et al., 2009) is a web application that constructs alignment, frequency, and weight matrices and displays them.

The generated matrices can be exported for use as input into pattern matching programs. The site can also generate the regular expression and IUPAC representations of the consensus motif. WebLogo displays a motif graphically so that sequence similarity can easily be visualized (Crooks et al., 2004).
It is important that the gathered TF binding sites are high quality since inaccuracies will produce a subpar matrix and consequently, poor motif matching performance (Medina-Rivera et al., 2010; Wittkop et al., 2010). Inaccuracies in TFBS information could stem from the imprecise nature of experimental approaches since gel shift, DNase footprinting, ChIP-chip, and ChIP-seq do not precisely identify binding sequences (Wittkop et al., 2010).
Two programs aim to analyze and optimize binding sequences. The utility 'matrix-quality' quantifies the ability of a matrix to distinguish background sequences and find functional binding sites in a genome (Medina-Rivera et al., 2010). It works by combining theoretical and empirical score distributions (Medina-Rivera et al., 2010). Another utility, MoRAine, goes one step further by shifting nucleotides around and takes the reverse complement of each TFBS sequence to try to improve the matrix (Wittkop et al., 2010) .

| Program | Platform | Web Address | Reference |
|---|---|---|---|
| D-MATRIX | Web | http://203.190.147.116/dmatrix | Sen et al., 2009 |
| matrix-quality | Web; Unix | http://rsat.ulb.ac.be/rsat | Medina-Rivera et al., 2010 |
| MoRAine | Web; Java | http://moraine.cebitec.uni-bielefeld.de | Wittkop et al., 2010 |
| WebLogo | Web; Unix | http://weblogo.berkeley.edu | Crooks et al., 2004 |

Table 3. Utility programs to manipulate transcription factor binding site sequences: construct and display frequency and weight matrices, generate regular expressions and IUPAC consensus patterns, and check and improve alignment quality. All the programs run inside a web browser (Platform = Web). Some of them can also be downloaded and executed locally on the user's computer running Unix or Unix-like operating system (Platform = Unix) or locally inside a Java virtual machine (Platform = Java). The Web Address column shows where the programs can be located or downloaded.

## 5. Pattern matching programs

Once a list of high quality TF binding sites is in hand, it can be fed into the pattern matching programs listed in Table 4 to find novel binding sites. All the listed programs are designed for prokaryotes or they have been tested with bacteria. This section briefly describes each motif matching program.
STAMP (Mahony & Benos, 2007) is not a true pattern matching program in that it does not scan genomes. Instead, it finds regulatory sequences deposited in motif databases that are most similar to a user-supplied set of binding sequences. It also performs multiple alignments on the supplied binding motifs and builds trees of the evolution of TF binding motifs.
MAST (Motif Alignment & Search Tool) (Bailey & Gribskov, 1998), a component of the MEME Suite, is one of the early programs that perform pattern matching on nucleotide (and

protein) sequences. The user can select one of the available genomes or upload a file containing up to one million nucleotides to search. The program requires an input file describing the matrix of the motif to search for.

| Program | Platform | Web Address | Reference |
|---|---|---|---|
| CRoSSeD | Web | http://ibiza.biw.kuleuven.be/crossed/webtool.html | Meysman et al., 2011 |
| EMBOSS > dreg | Web; Unix | http://emboss.sourceforge.net | Rice et al., 2000 |
| dscan | Web | http://bayesweb.wadsworth.org/cgi-bin/dscan.pl | Thompson et al., 2005 |
| FITBAR | Web | http://archaea.u-psud.fr/fitbar | Oberto, 2010 |
| iMotifs | Mac | http://wiki.github.com/mz2/imotifs | Piipari et al., 2010 |
| MAST | Web; Unix | http://meme.sdsc.edu/meme/mast-intro.html | Bailey & Gribskov, 1998 |
| Motif Locator | Web | http://www.cmbl.uga.edu/software.html | Mrazek et al., 2008 |
| MyPattern Finder | Web | http://www.nii.ac.in/~deepak/RegAnalyst | Sharma et al., 2009 |
| PatScan | Unix | http://ftp.mcs.anl.gov/pub/Genomics/PatScan | Dsouza et al., 1997 |
| Pattern Locator | Web; Unix | http://www.cmbl.uga.edu/software.html | Mrazek & Xie, 2006 |
| PhyloScan | Web | http://bayesweb.wadsworth.org/phyloscan | Palumbo & Newberg, 2010 |
| PredictRegulon | Web | http://www.cdfd.org.in/predictregulon | Yellaboina et al., 2004 |
| RegPredict | Web | http://regpredict.lbl.gov | Novichkov et al., 2010b |
| RSAT | Web | http://rsat.ulb.ac.be/rsat | Thomas-Chollier et al., 2008 |
| SITECON | Web | http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon | Oshchepkov et al., 2004 |
| STAMP | Web | http://www.benoslab.pitt.edu/stamp | Mahony & Benos, 2007 |
| Virtual Footprint | Web | http://www.prodoric.de/vfp | Munch et al., 2005 |

Table 4. Programs that can scan prokaryote genomes for transcription factor binding sites. All these programs can run over the web inside a browser (Platform = Web) except iMotifs, a MacOS X only application, and PatScan, a program that must be downloaded and run locally on Unix or Unix-like systems. Three of the web applications—dreg, MAST, and Pattern Locator—can also be downloaded and execute on Unix or Unix-like systems. The Web Address column shows where a program can be run or downloaded.

PatScan (Dsouza et al., 1997) is another early motif matching program. Even though it is designed to search protein sequences for motifs and nucleotide sequences for hairpins,

pseudoknots, repeats, and other secondary structures, it could be used to search genomic DNA for TFBSs. Mismatches, insertions, and deletions are allowed. It runs on Unix systems only. A web version seems to be no longer available.

The program 'dreg' searches one or more sequences for a given motif described by a regular expression (Rice et al., 2000). It is one of the hundreds of tools comprising EMBOSS (European Molecular Biology Open Software Suite). EMBOSS' mission is to provide a place to bring together the rapid increase in the number of complete genomes and new sequence analysis software and make them publicly available as a suite. The tools perform sequence alignment, database searching with sequence patterns, nucleotide sequence patterns (CpG islands, repeats, etc.), and more.

To address the usability issue associated with PatScan and dreg, Pattern Locator was created (Mrazek & Xie, 2006). Its purpose is to find short sequence patterns in complete genomes. The input string uses a special syntax or it can be specified using the IUPAC alphabet. The flexible syntax allows the following to be specified: direct and inverted repeats, maximum number of mismatches allowed, direct or complementary DNA strand to search, and gaps.

Many other programs do not require motifs to be supplied in a special syntax. Motif Locator (Mrazek et al., 2008) takes a set of binding sequences, turns it into a matrix, and uses the matrix to search a genome for instances of the motif. MyPatternFinder (Sharma et al., 2009) finds exact or approximate occurrences of a motif from a selection of over 600 complete genomes using an exact search method and an alignment technique. Insertions and deletions are allowed. The program 'dscan' (Thompson et al., 2005) scans genome databases for statistically significant sites similar to the given motif. Two databases of *E. coli* and *Rhodopseudomonas palustris* intergenic regions are provided.

PredictRegulon (Yellaboina et al., 2004) is a web application that scans a prokaryote genome for potential target genes of a TF. The user picks a bacterial genome from a list of over 110 and supplies a set of aligned binding site sequences for the transcription factor. The program then scans the upstream sequences of all the genes in the selected genome, calculates a score for a potential binding site in each promoter, and outputs the site if the score is above the threshold cutoff value, which is taken to be the lowest score in the input sequence set. The output includes the binding site sequence, the name and description of the gene, and operon context and detailed information on the gene.

FITBAR (Fast Investigation Tool for Bacterial and Archaeal Regulons) (Oberto, 2010) is a matrix search program that scans whole Bacteria and Archaea genomes retrieved from the National Center for Biotechnology Information repository to discover sets of genes regulated by TFs. Unlike most other genome matching programs such as PredictRegulon, FITBAR does not find matches by arbitrary score cutoff values. It uses the log-odds and entropy-weighted search algorithms and Compound Importance Sampling (CIS) and Local Markov Method (LMM) to calculate the statistical significance of the predicted motifs (p-values). Aligned TFBS sequences can be supplied, or one of the 200 known prokaryotic matrices can be selected. Results are listed, along with a graphical depiction of the motif sequence location and the surrounding genes.

Like EMBOSS, RSAT (Regulatory Sequence Analysis Tools) (Thomas-Chollier et al., 2008) contains a collection of tools to analyze *cis*-acting regulatory elements in the noncoding sequences from over 600 genomes. The tools perform both pattern matching (and pattern discovery) and return information on individual genes, such as orthologs and DNA sequences. Namely, the five pattern matching programs—*dna-pattern*, *genome-scale-dna-*

*pattern*, *matrix-scan*, *patser*, and *genome-scale-patser*—allow one to search entire genomes or a set of sequences for occurrences of a motif represented as a regular expression, an IUPAC string, or a position weight matrix. Various statistical background models are available to allow the significance of the matches to be evaluated.

Virtual Footprint (Munch et al., 2005) is an online interactive environment to search and analyze TFBS, gapped or ungapped, in bacterial genomes. The TFBS pattern to search can be picked from a list of pre-defined matrix motifs, or a set of sequences, a regular expression, or IUPAC codes can be supplied. A match is assigned to a gene if possible and the genomic context is provided. The program can check if a match also occurs in the regulatory region of orthologous genes.

Like Virtual Footprint, iMotifs (Piipari et al., 2010) provides an integrated environment to visualize, analyze, and annotate sequence motifs. However, it does not run over the web. It is a Java-based application that runs on MacOS X only.

More advanced pattern matching programs incorporate the use of cross-species conservations during their genome search to enrich the predicted sites. Comparative genomics approach is predicated on the idea that TFs from related organisms regulate genes that tend to be conserved (Novichkov et al., 2010b). Presence of similar TFBSs upstream of orthologous genes increases the probability that the sites are functional binding sites (Novichkov et al., 2010b).

PhyloScan (Carmack et al., 2007; Palumbo & Newberg, 2010) is a web program that screens candidate sequences by using (1) aligned or unaligned sequence data from multiple species, even evolutionarily distant ones, (2) multiple sites within an intergenic region, and (3) q-values to predict more functional TFBSs, even weak ones, in a genome. The use of q-values is in contrast to conventional motif matching programs, which either score a candidate binding site against a training set of TFBS or evaluate the statistical significance of the candidate binding site using p-value.

Another program that takes the comparative genomics approach is RegPredict (Novichkov et al., 2010b), a web site that provides a visual environment for the discovery of genes regulated by a TF in prokaryotes. The site contains a large collection of known TFBS motifs gathered from the RegPrecise, RegTransBase, and RegulonDB databases and genomic sequences of major taxonomic groups of Bacteria. Any of the motifs can be selected, or the user can upload a set of aligned binding site sequences, and RegPredict will scan for the motif in up to 15 genomes simultaneously. (If the regulatory motif is not known, RegPredict can predict one *de novo* from user-supplied coregulated genes.) Candidate genes are grouped into different clusters based on the degree of conservation of regulatory interactions and then presented in a multi-pane user interface, along with the genomic context and gene function information, for the user to analyze.

Other advanced motif matching programs use DNA structure information to increase their performance. A factor that contributes to the specificity of the interaction between a TF and its binding site is the local conformation of the DNA site (Oshchepkov et al., 2004). Even though a TF often regulates multiple genes and the binding sites in the promoters of these genes show variations, certain conformational and physicochemical properties are conserved among these sites so that the TF can recognize the sites (Oshchepkov et al., 2004). Thus, these context-dependent TFBS properties can be used to improve the predictions of genes controlled by a TF (Oshchepkov et al., 2004).

SITECON (Oshchepkov et al., 2004) is a web application that can analyze and report 38 properties—major groove depth, bend, entropy change, to name a few—in a given set of DNA binding site sequences, and optionally, find binding sites in one or more DNA sequences using those properties.
CRoSSeD (Conditional Random fields of Smoothed Structural Data) (Meysman et al., 2011) is another program that leverages structure information. Specifically, it uses 12 structural scales, such as protein-induced deformability and stabilization energy, that are presumably relevant to binding site recognition in prokaryotes. (Scales are experimentally determined models for approximating regional DNA structure based on di- or trinucleotides.) Some of the novel binding sites found by CRoSSeD had low sequence similarity. A check with the literature and database indicated that they may be true binding sites. This shows that searching for binding sites based on structure information is a viable approach since these binding sites, with their weak motif, may be missed by traditional pattern matching programs.

## 6. EnvZ/OmpR regulon prediction

Our lab is currently using genetic, biochemical, and bioinformatics approaches to determine the set of genes regulated by OmpR in *E. coli*. A microarray experiment showed that the expression levels of 125 genes were significantly affected in an EnvZ-null background (Oshima et al., 2002). To help identify the genes that are directly modulated by OmpR, we searched the RegulonDB databank and found 23 OmpR binding sites for 11 genes, as listed in Table 5.
Using all of the OmpR binding sequences except *ecnB*'s (since it is not 20-bp long) as input, the pattern matching program Motif Locator detected 12,314 matches in the intergenic regions of the *E. coli* K12 genome. Since *E. coli* has over 4,200 genes (Blattner et al., 1997), the results clearly contained many false positives.

| Gene | OmpR Binding Site | Gene | OmpR Binding Site |
|---|---|---|---|
| *bolA* | AACCTAAATATTTGTTGTTA | *micF* | CGAATATGATACTAAAACTT |
| *nmpC* | AACTTACATCTTGAAATAAT | *micF* | TTAAGATGTTTCATTTATCG |
| *ompF* | TTTACTTTTGGTTACATATT | *micF* | TATAGATGTTTCAAAATGTA |
| *ompF* | TTTTCTTTTTGAAACCAAAT | *ompC* | TTTACATTTTGAAACATCTA |
| *ompF* | CTTTATCTTTGTAGCACTTT | *ompC* | AGCGATAAATGAAACATCTT |
| *ompF* | GTTACGGAATATTACATTGC | *ompC* | AAAAGTTTTAGTATCATATT |
| *csgD* | TACATTTAGTTACATGTTTA | *fadL* | GAGCCAGAAAACCCTGTTTA |
| *tppB* | GTAACAGATTATTACAAAGG | *fadL* | TTAGATCATATTTGAAAAAA |
| *flhD* | AAAAATCTTAGATAAGTGTA | *fadL* | ACGTAACATAGTTTGTATAA |
| *flhD* | GGGCATTATCTGAACATAAA | *fadL* | AAATCACACTTAAAAATGAT |
| *omrA* | CACACCTCGTTGCATTTCCC | *ecnB* | AACATAAATAACAT |
| *omrB* | AACCTTTGGTTACACTTTGC | | |

Table 5. List of known OmpR binding sites and the corresponding genes. The list was compiled using RegulonDB.

To increase the specificity and reduce the number of matches returned, we picked 10 binding sites from five genes: *ompF*, *ompC*, *tppB*, *csgD*, and *fadL*. See Table 6. These sequences

were chosen because each contains two direct repeats of the consensus motif GTTACANNNN, which is derived from extensive studies on interactions between OmpR and *ompF* and *ompC* promoters (Harlocker et al., 1995; Yoshida et al., 2006). Note that we adjusted the alignment of the *csgD* and *fadL* binding sequences to better fit the consensus model. The *csgD* and *fadL* sequences from RegulonDB shown in Table 5 span from –57 to –38 and from +58 to +77 relative to the transcriptional start site, whereas the adjusted ones span from -59 to -40 and from 50 to 69, respectively.

Pattern matching analysis of the 10 sequences listed in Table 6 using Motif Locator found 110 matches, five of which were among the 125 genes affected in the microarray experiment: *ompF*, *flgL*, *ompC*, *rpoE*, and *cysC*. The same set of 10 sequences was fed into another pattern matching program, Virtual Footprint, which predicted 32 genes modulated by OmpR. Four genes were the same as those identified in the microarray data: *ompF*, *ydgR*, *ompC*, and *ygjU*. Only two genes were found by both Motif Locator and Virtual Footprint, *ompF* and *ompC*, showing that different programs return different results.

| Gene | OmpR Binding Site |
|---|---|
| *ompF* | TTTACTTTTGGTTACATATT |
| *ompF* | TTTTCTTTTTGAAACCAAAT |
| *ompF* | CTTTATCTTTGTAGCACTTT |
| *ompF* | GTTACGGAATATTACATTGC |
| *ompC* | TTTACATTTTGAAACATCTA |
| *ompC* | AGCGATAAATGAAACATCTT |
| *ompC* | AAAAGTTTTAGTATCATATT |
| *tppB* | GTAACAGATTATTACAAAGG |
| *csgD* | GTTACATTTAGTTACATGTT |
| *fadL* | GTTACAGCACGTAACATAGT |

Table 6. OmpR binding sequences that contain direct repeats of the GTTACANNNN consensus motif, where N denotes any nucleotide.

The degenerate OmpR binding motif makes identification of new regulon member difficult. When the set of 10 sequences in Table 6 was used as input, Motif Locator predicted only half of the 12 known OmpR regulated genes: *bolA*, *ompF*, *csgD*, *micF*, *ompC*, and *fadL*, whereas Virtual Footprint returned four: *ompF*, *micF*, *ompC*, and *fadL*. This observation suggests that the run was too specific and more novel genes remain to be discovered. To find them, one can try different sets of input sequences, run other pattern matching programs, or make use of comparative genomics or published OmpR crystal structures (Kondo et al., 1997; Martínez-Hackert & Stock, 1997).

## 7. Conclusion

Like our own experience of using bioinformatics tools to study the OmpR regulon illustrates, comparative studies on the performance of motif discovery and matching programs found no single program works well on all data sets (MacIsaac & Fraenkel, 2006). In particular, a benchmark of four motif matching programs—RSA Tools, PRODORIC Virtual Footprint, RegPredict, and FITBAR—for their ability to discover potential binding sites for the transcriptional regulator NagC involved in N-acetylglucosamine metabolism in

the *Escherichia coli* K12 MG1655 genome found that some tools uncover sites that others have missed (Oberto, 2010). Therefore, it is recommended that multiple tools be used instead of just one and the output from multiple programs be combined and compared in order to improve accuracy and gain confidence in the results (Das & Dai, 2007; Mrazek, 2009).

The easiest way to run the pattern matching programs—and other bioinformatics tools—is over the web inside a browser. However, in order to help keep the load on the web servers hosting these programs to a low level, some sites put a limit on the complexity of the jobs submitted. If a web site places such restriction, a desktop version of the program is usually provided for users to download and install or compile on their local computer. Many of the desktop programs run in a Java environment or on Unix or Unix-like system, such as Linux. Some Unix programs can run on Windows if the Linux-like environment Cygwin is set up first. However, it should be noted that setting up the required runtime environment and installing or compiling these programs take considerable effort and computer expertise. Also be aware that some desktop programs, especially those that run on Unix, are run from the command line; there is no graphical user interface.

The identification of a TF's binding motif and the identification of new target genes are difficult to do experimentally and computationally (Pavesi et al., 2004) because we do not completely understand the biology of gene regulation (Das & Dai, 2007). But it is hoped that the information in this chapter will make the task of pattern matching easier for microbiologists and other researchers.

## 8. Acknowledgment

This work was supported by grant GM048167 from the National Institutes of Health.

## 9. References

Bailey, T. L. & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, Vol. 14, No. 1, pp. 48-54.

Bauer, A. L., Hlavacek, W. S., Unkefer, P. J., & Mu, F. (2010). Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Comput Biol*, Vol. 6, No. 11, pp. e1001007.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, Vol. 277, No. 5331, pp. 1453-62.

Bulyk, M. L. (2003). Computational prediction of transcription-factor binding site locations. *Genome Biol*, Vol. 5, No. 1, pp. 201.

Carmack, C. S., McCue, L. A., Newberg, L. A., & Lawrence, C. E. (2007). PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms Mol Biol*, Vol. 2, pp. 1.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, Vol. 14, No. 6, pp. 1188-90.

Das, M. K. & Dai, H. K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*, Vol. 8 Suppl 7, pp. S21.

D'Haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nat Biotechnol*, Vol. 24, No. 8, pp. 959-61.

D'Haeseleer, P. (2006b). What are DNA sequence motifs? *Nat Biotechnol*, Vol. 24, No. 4, pp. 423-5.

Dsouza, M., Larsen, N., & Overbeek, R. (1997). Searching for patterns in genomic data. *Trends in Genetics : TIG*, Vol. 13, No. 12, pp. 497-8.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., Garcia-Sotelo, J. S., Lopez-Fuentes, A., Porron-Sotelo, L., Alquicira-Hernandez, S., Medina-Rivera, A., Martinez-Flores, I., Alquicira-Hernandez, K., Martinez-Adame, R., Bonavides-Martinez, C., Miranda-Rios, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., & Collado-Vides, J. (2011). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research*, Vol. 39, No. Database issue, pp. D98-105.

Grote, A., Klein, J., Retter, I., Haddad, I., Behling, S., Bunk, B., Biegler, I., Yarmolinetz, S., Jahn, D., & Munch, R. (2009). PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Research*, Vol. 37, No. Database issue, pp. D61-5.

Harlocker, S. L., Bergstrom, L., & Inouye, M. (1995). Tandem binding of six OmpR proteins to the ompF upstream regulatory sequence of Escherichia coli. *The Journal of Biological Chemistry*, Vol. 270, No. 45, pp. 26849-56.

Homann, O. R. & Johnson, A. D. (2010). MochiView: versatile software for genome browsing and DNA motif analysis. *BMC Biol*, Vol. 8, pp. 49.

Hu, J., Li, B., & Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, Vol. 33, No. 15, pp. 4899-913.

Kondo, H., Nakagawa, A., Nishihira, J., Nishimura, Y., Mizuno, T., & Tanaka, I. (1997). Escherichia coli positive regulator OmpR has a large loop structure at the putative RNA polymerase interaction site. *Nature Structural Biology*, Vol. 4, No. 1, pp. 28-31.

Ladunga, I. (2010). An overview of the computational analyses and discovery of transcription factor binding sites. *Methods Mol Biol*, Vol. 674, pp. 1-22.

MacIsaac, K. D. & Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, Vol. 2, No. 4, pp. e36.

Mahony, S. & Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, Vol. 35, No. Web Server issue, pp. W253-8.

Martinez-Hackert, E. & Stock, A. M. (1997). The DNA-binding domain of OmpR: crystal structures of a winged helix transcription factor. *Structure*, Vol. 5, No. 1, pp. 109-24.

Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., & van Helden, J. (2010). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res*, Vol. 39, No. 3, pp. 808-24.

Meysman, P., Dang, T. H., Laukens, K., De Smet, R., Wu, Y., Marchal, K., & Engelen, K. (2011). Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. *Nucleic Acids Res*, Vol. 39, No. 2, pp. e6.

Mrazek, J. (2009). Finding sequence motifs in prokaryotic genomes--a brief practical guide for a microbiologist. *Brief Bioinform*, Vol. 10, No. 5, pp. 525-36.

Mrazek, J. & Xie, S. (2006). Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. *Bioinformatics*, Vol. 22, No. 24, pp. 3099-100.

Mrazek, J., Xie, S., Guo, X., & Srivastava, A. (2008). AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes. *Bioinformatics*, Vol. 24, No. 8, pp. 1041-8.

Munch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M., & Jahn, D. (2005). Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, Vol. 21, No. 22, pp. 4187-9.

Novichkov, P. S., Laikova, O. N., Novichkova, E. S., Gelfand, M. S., Arkin, A. P., Dubchak, I., & Rodionov, D. A. (2010a). RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Research*, Vol. 38, No. Database issue, pp. D111-8.

Novichkov, P. S., Rodionov, D. A., Stavrovskaya, E. D., Novichkova, E. S., Kazakov, A. E., Gelfand, M. S., Arkin, A. P., Mironov, A. A., & Dubchak, I. (2010b). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic acids research*, Vol. 38, No. Web Server issue, pp. W299-307.

Oberto, J. (2010). FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics*, Vol. 11, pp. 554.

Oshchepkov, D. Y., Vityaev, E. E., Grigorovich, D. A., Ignatieva, E. V., & Khlebodarova, T. M. (2004). SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucleic Acids Res*, Vol. 32, No. Web Server issue, pp. W208-12.

Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B. L., Mori, H., & Mizuno, T. (2002). Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. *Molecular Microbiology*, Vol. 46, No. 1, pp. 281-91.

Palumbo, M. J. & Newberg, L. A. (2010). Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. *Nucleic Acids Research*, Vol. 38, No. Web Server issue, pp. W268-74.

Pavesi, G., Mauri, G., & Pesole, G. (2004). In silico representation and discovery of transcription factor binding sites. *Brief Bioinform*, Vol. 5, No. 3, pp. 217-36.

Piipari, M., Down, T. A., Saini, H., Enright, A., & Hubbard, T. J. (2010). iMotifs: an integrated sequence motif visualization and analysis environment. *Bioinformatics*, Vol. 26, No. 6, pp. 843-4.

Quest, D., Dempsey, K., Shafiullah, M., Bastola, D., & Ali, H. (2008). MTAP: the motif tool assessment platform. *BMC Bioinformatics*, Vol. 9 Suppl 9, pp. S6.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, Vol. 16, No. 6, pp. 276-7.

Schug, J. (2008). Using TESS to predict transcription factor binding sites in DNA sequence. *Curr Protoc Bioinformatics*, Vol. Chapter 2, pp. Unit 2 6.

Sen, N., Mishra, M., Khan, F., Meena, A., & Sharma, A. (2009). D-MATRIX: a web tool for constructing weight matrix of conserved DNA motifs. *Bioinformation*, Vol. 3, No. 10, pp. 415-8.

Sharma, D., Mohanty, D., & Surolia, A. (2009). RegAnalyst: a web interface for the analysis of regulatory motifs, networks and pathways. *Nucleic Acids Res*, Vol. 37, No. Web Server issue, pp. W193-201.

Stormo, G. D. (2010). Motif discovery using expectation maximization and Gibbs' sampling. *Methods Mol Biol*, Vol. 674, pp. 85-95.

Tan, K., McCue, L. A., & Stormo, G. D. (2005). Making connections between novel transcription factors and their DNA motifs. *Genome Res*, Vol. 15, No. 2, pp. 312-20.

Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., & van Helden, J. (2008). RSAT: regulatory sequence analysis tools. *Nucleic acids research*, Vol. 36, No. Web Server issue, pp. W119-27.

Thompson, W., Conlan, S., McCue, L. A., & Lawrence, C. E. (2007). Using the Gibbs Motif Sampler for phylogenetic footprinting. *Methods Mol Biol*, Vol. 395, pp. 403-24.

Thompson, W., McCue, L. A., & Lawrence, C. E. (2005). Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. *Curr Protoc Bioinformatics*, Vol. Chapter 2, pp. Unit 2 8.

Wei, W. & Yu, X. D. (2007). Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. *Genomics Proteomics Bioinformatics*, Vol. 5, No. 2, pp. 131-42.

Wittkop, T., Rahmann, S., & Baumbach, J. (2010). Efficient online transcription factor binding site adjustment by integrating transitive graph projection with MoRAine 2.0. *J Integr Bioinform*, Vol. 7, No. 3, pp. 1-11.

Yanover, C., Singh, M., & Zaslavsky, E. (2009). M are better than one: an ensemble-based motif finder and its application to regulatory element prediction. *Bioinformatics*, Vol. 25, No. 7, pp. 868-74.

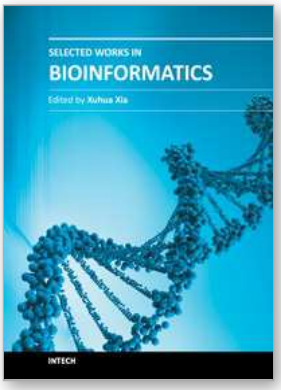
Yellaboina, S., Seshadri, J., Kumar, M. S., & Ranjan, A. (2004). PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic acids research*, Vol. 32, No. Web Server issue, pp. W318-20.

Yoshida, T., Qin, L., Egger, L. A., & Inouye, M. (2006). Transcription regulation of ompF and ompC by a single transcription factor, OmpR. *The Journal of Biological Chemistry*, Vol. 281, No. 25, pp. 17114-23.

Zaslavsky, E. & Singh, M. (2006). A combinatorial optimization approach for diverse motif finding applications. *Algorithms Mol Biol*, Vol. 1, pp. 13.

Zhang, S., Xu, M., Li, S., & Su, Z. (2009). Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res*, Vol. 37, No. 10, pp. e72.

Zhou, Q. & Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, Vol. 20, No. 6, pp. 909-16.




**Selected Works in Bioinformatics**

Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
**Publisher** InTech
**Published online** 19, October, 2011
**Published in print edition** October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Phu Vuong and Rajeev Misra (2011). Guide to Genome-Wide Bacterial Transcription Factor Binding Site Prediction Using OmpR as Model, Selected Works in Bioinformatics, Dr. Xuhua Xia (Ed.), ISBN: 978-953-307-281-4, InTech, Available from: http://www.intechopen.com/books/selected-works-in-bioinformatics/guide-to-genome-wide-bacterial-transcription-factor-binding-site-prediction-using-ompr-as-model

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821