8

Prediction and Analysis of Gene Regulatory Networks in Prokaryotic Genomes

Richard Münch, Johannes Klein and Dieter Jahn
Institute of Microbiology, Technische Universität Braunschweig, Braunschweig Germany

1. Introduction

The availability of over 1500 completely sequenced and annotated prokaryotic genomes offers a variety of comparative and predictive approaches on genome-scale. The results of such analyses strongly rely on the quality of the employed data and the computational strategy of their interpretation. Today, comparative genomics allows for the quick and accurate assignment of genes and often their corresponding functions. The resulting list of classified genes provides information about the overall genomic arrangement, of metabolic capabilities, general and unique cellular functions, however, almost nothing about the underlying complex regulatory networks. Transcriptional regulation of gene expression is a central part of these networks in all organisms. It determines the actual RNA, protein and as a consequence metabolite composition of a cell. Moreover, it allows cells to adapt these parameters in response to changing environmental conditions. An integral part of transcriptional regulation is the specific interaction of transcription factors (TFs) with their corresponding DNA targets, the transcription factor binding sites (TFBSs) or motifs. Recent advances in extensive data mining using various high-throughput techniques provided first insights into the complex regulatory networks and their interconnections. However, the computational prediction of regulatory interactions in the promoter regions of identified genes remains to be difficult. Consequently, there is a high demand for the *in silico* identification and analysis of involved regulatory DNA sequences and the development of software tools for the accurate prediction of TFBSs.

In this chapter we focus on methods for the prediction of TFBSs in whole prokaryotic genomes (regulons). Although, many studies were sucessfully performed in eukaryotes they are often not transferable to the special features of bacterial gene regulation. In particular the prokaryotic genome organization concerning clusters of co-transcribed polycistronic genes, the lack of introns and the shortness of promoter sequences necessitates adapted computational approaches. Besides the genomic structure there are also differences in the regulatory control logic. Prokaryotic promoters often possess one or few regulatory interactions while the repertoire of regulators consists of only a couple of global TFs but many local TFs (Price et al., 2008). On the other hand, eukaryotic promoters and enhancers involve the concerted binding of multiple regulators, so called cis-regulatory modules (CRMs) or composite elements (Loo & Marynen, 2009). Many excellent reviews in the field prokaryotic gene regulation were recently published with focus on the broad spectrum of approaches for the experimental and theoretical reconstruction of gene regulatory networks and their

interspecies transfer (Baumbach, 2010; Rodionov, 2007; van Hijum et al., 2009; Zhou & Yang, 2006). Here, we focus on practical aspects how to detect new members of a regulon for genes or genomes of interest. We will summarize useful bioinformatics databases, methods and algorithms available for unraveling bacterial gene regulatory networks from whole genome sequences. Finally, we want to indicate the limitations and technical problems of such approaches and give a survey on recent improvements in this field.

2. Strategies for the prediction of transcripion factor binding sites

Basically, today exist at least two general approaches to recognize regulatory sequence patterns. One challenging approach called **pattern discovery** relies on a statistical overrepresentation of DNA sequence motifs present in promoters of structurally and funktionally related or co-regulated genes. In that case it is a *de-novo* prediction where the binding site and the corresponding regulator are unknown. The list of investigated genes can be derived from clusters of co-expressed genes available in microarray experiments, from ChIP-on-chip experiments or from orthologous genes of related organisms. In the latter case this method is called phylogenetic footprinting (McCue et al., 2001). Pattern discovery algorithms are top-down approaches that use various learning principles with different degrees of performance (Sandve et al., 2007; Su et al., 2010; Tompa et al., 2005). The advantage of this method is the detection of potential regulatory DNA sequences even if there is little known about the corresponding regulation. A recent study in prokaryotes appling a pattern discovery approach revealed that the predicted patterns matched up to 81% of known individual TFBSs (Zhang et al., 2009). However, this approach has limitted value in getting a clue about what specific regulator is involved in a predicted TFBS.

An alternative approach on which we focus in this chapter is called **pattern matching**. It makes use of prior knowledge in form of a predetermined pattern that can be assigned to a specific regulator. The pattern is usually build based on a profile of known TFBSs for which experimental evidence is available (Fig. 1 A). Using this set of DNA sequences a probabilistic model describing the pattern degeneracy is constructed. Application of the model on a given sequence results in a score for the likelihood that the investigated sequence belongs to the same sequence family. The application of pattern matching involves the availability of a reliable training set of TFBSs. For that purpose, several specalized databases provide collections and patterns of prokaryotic TBFSs supplemented with various related information like promoter and operon structures. A limited list of important data sources is shown in table 1.

In the following examples a data set of 40 experimentally proven TFBSs from the anaerobic regulator Anr of *Pseudomonas aeruginosa* is used (Trunk et al., 2010). There are different ways of pattern representation. Traditionally, the usage of IUPAC code for base ambiguities is a straightforward way to describe a binding motif (NC-IUB, 1985). In this approach, combinations of certain bases are assigned to an extended alphabet of specific letters (Fig. 1 B). IUPAC code can be easily converted into a regular expression (Fig. 1 C). A regular expression is a formal language for pattern matching, that can be used to scan for ambiguous IUPAC strings in order to predict new TFBSs (Betel & Hogue, 2002). (Fig. 1 B). Allthough the IUPAC letter code is very concise and still widely used among biologists it does not describe a proper weighting of bases. Additionally, the majority rules how to generate a consensus sequences are to some extent arbitrary (Day & McMorris, 1992). However, in the case that the training set consists of only a few sequences the usage of IUPAC code can still make sense.

Name	Year	Data content	URL	References
CoryneRegNet	2006	*Coynebacerium* TFBSs, regulatory networks, predictions	http://www.coryneregnet.de	Baumbach et al. (2009)
DBTBS	2001	*B. subtilis* TFBSs, operons, predictions	http://dbtbs.hgc.jp	Sierro et al. (2008)
DPInteract	1998	*E. coli* TFBSs, PWMs	http://arep.med.harvard.edu /dpinteract	Robison et al. (1998)
PRODORIC	2003	prokaryotic TFBSs, PWMs, promoters, expression data	http://www.prodoric.de	Grote et al. (2009)
PromEC	2001	*E. coli* promoters	http://margalit.huji.ac.il /promec	Hershberg et al. (2001)
RegPrecise	2010	predicted TFBSs	http://regprecise.lbl.go	Novichkov et al. (2010)
RegTransBase	2007	prokaryotic TFBSs, PWMs	http://regtransbase.lbl.gov	Kazakov et al. (2007)
RegulonDB	1998	*E. coli* TFBSs, PWMs, operons,	http://regulondb.ccg.unam.mx	Gama-Castro et al. (2011)
Tractor_DB	2004	predicted TFBSs of γ-proteobacteria	http://www.tractor.lncc.br	Pérez et al. (2007)

Table 1. List of important public databases about bacterial gene regulation. The table shows the name, year of establishment, data content, the internet address and the latest reference of the respective database.

A more accurate description of a binding pattern is achieved by probabilistic models like a frequency matrix (or alignment matrix) (Staden, 1984). Instead of considering only the most common bases at each position a matrix comprises the frequencies for each nucleotide at each position (Fig. 1 D). Based on frequency matrices many models for the calculation of weights were proposed. Such a model is broadly called position weight matrix (PWM) or position specific scoring matrix (PSSM). PWMs can be considered as simplified profile hidden Markov models (HMM) that do not allow insertion and deletion states (Durbin et al., 1998). Formally, a PWM is an array M of weights w where each column corresponds to the position of the TFBS motif of the length l and each row represents the letter of the sequence alphabet $\mathcal{A}$. In case of DNA $\mathcal{A} \in \{A, C, G, T\}$ (equation 1).

$$M = \begin{vmatrix} w_{A,1} & w_{A,2} & \cdots & w_{A,l} \\ w_{C,1} & w_{C,2} & \cdots & w_{C,l} \\ w_{G,1} & w_{G,2} & \cdots & w_{G,l} \\ w_{T,1} & w_{T,2} & \cdots & w_{T,l} \end{vmatrix} \quad (1)$$

Many very related examples for the calculation of individual weights were proposed in the literaure (Berg & von Hippel, 1987; Fickett, 1996; Schneider et al., 1986; Staden, 1984; Stormo, 2000). The information theoretical approach and modifications of it ((Schneider et al., 1986)) are widely used and some of the most successful methods for both the modeling and the prediction of potential TFBSs. Information is a measure of uncertainty which means that

a highly conserved position with the exclusive occurence of one specific nucleotide gets the highest information value of 2 bits. In other words there is a maximum certainty of finding this nucleotide at this position. In contrast, an information value of 0 bits represents a highly degenerated position and the highest uncertainty of finding a specific nucleotide. The information vector $R(l)$ represents the total information content of a profile of aligned sequences at the position l with $f(b,l)$ indicating the frequency of the base b at position l.

$$R(l) = 2 + \sum_{b=A}^{T} f(b,l) \log_2 f(b,l) \tag{2}$$

An information PWM $m(b,l)$ is generated by multiplying the base frequencies $f(b,l)$ with the total information content $R(l)$ (Fig. 1 E).

$$m(b,l) = f(b,l) \cdot R(l) \tag{3}$$

For pattern matching applications a PWM is used by summing up the corresponding weights of a candidate sequence to a score. Afterwards, these scores are compared to a predefined cut-off (or threshold) to filter out potential predictions. The derived score is often correlated to the binding affinity of a TF thus the information score can be interpreted as an rough estimate to the specific bindung energy. However, this is only possible under the simplifying assumption that each position of a pattern contributes independently to the TF-TFBS interaction. This additivity assumption is controversially discussed but is was shown that it is in fact a reasonable approximation (Benos et al., 2002). The graphical representation of an information PWM is called sequence logo (Schneider & Stephens, 1990). In a sequence logo each PWM weight is equivalent to the individual letter size so the total height of the stack of letters represents the information content $R(l)$ at this position. Sequence logos allow an illustrative visualization of the sequence conservation and binding preference of a regulator (Fig. 1 F).

3. Statistical significance of pattern matching

Regulatory sequences are commonly short (usually 6-18 bp), the sample size of experimentally proven sites is often limited and in many cases the observed level of sequence conservation is low. Consequently, the genome-wide statistically occurance frequency of derived patterns is often unrealistically high. In such cases, searches generally generate increasing numbers of false-predictions the lower the threshold score is set. This is demonstrated in Fig. 2 showing the score distributions of true and false predictions of a genome wide search in *P. aeruginosa* using the PWM of the Anr regulator (Fig. 1 E). In the shown example matches in coding regions were considered as false-predictions (false-positives) and matches that are part of the training set were naturally ranked as true-predictions (true-positives). Score distributions are also important indicators to evaluate the predictive capacity of a PWM (Medina-Rivera et al., 2011).

In order to improve the predictive power of pattern matching, commonly a cut-off score is set in a way, that improves the ratio of true- and false-predictions. However, thereby the total number of hits will still contain to some extent false-positives while some true matches become lost (false-negatives). From this it follows that matches of TFBS predictions can not be classified in a binary manner like a dignostic test, since true-positives and false-positives are always coexisting. Alternatively, they can be grouped into a classification schema consisting

A) Excerpt of 40 Sample sequences (training set)

1	T	T	G	A	T	T	C	C	G	G	T	C	A	A
2	T	T	G	A	C	T	T	T	C	A	T	C	A	A
3	T	T	G	A	T	T	G	C	C	A	T	C	A	A
4	T	T	G	A	C	C	G	G	A	A	T	C	A	A
⋮								⋮						
40	T	T	G	A	T	G	T	C	G	A	T	C	A	A

B) IUPAC consensus

Y	T	G	H	Y	N	B	N	B	V	K	C	A	R

C) Regular Expression

[CT]TG[ACT][CT][ACGT][CGT][ACGT][CGT][ACG][TG]CA[AG]

D) Frequency Matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	2	0	0	29	0	6	0	10	6	20	0	0	40	35
C	5	0	0	7	17	17	15	10	15	3	1	40	0	0
G	0	0	40	0	3	10	9	18	16	17	2	0	0	5
T	33	40	0	4	20	7	16	2	3	0	37	0	0	0

E) Position Weight Matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	0.06	0.00	0.00	0.65	0.00	0.02	0.00	0.07	0.04	0.35	0.00	0.00	2.00	1.27
C	0.15	0.00	0.00	0.16	0.30	0.05	0.17	0.07	0.09	0.05	0.04	2.00	0.00	0.00
G	0.00	0.00	2.00	0.00	0.05	0.03	0.10	0.12	0.10	0.30	0.08	0.00	0.00	0.18
T	0.97	2.00	0.00	0.09	0.35	0.02	0.18	0.01	0.02	0.00	1.43	0.00	0.00	0.00

F) Sequence Logo

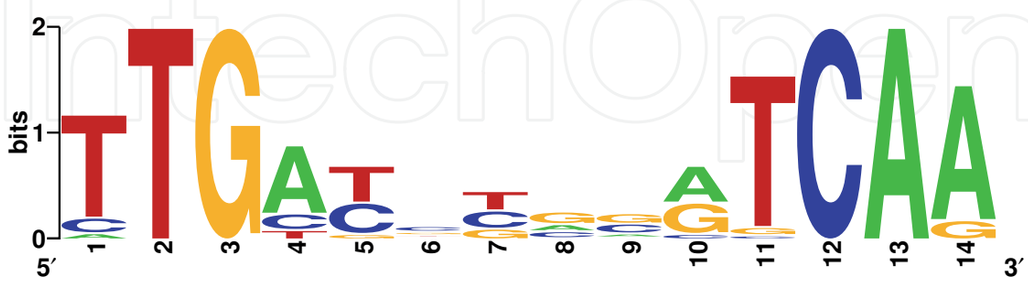


Fig. 1. Various pattern representations for a taining set 40 Anr binding sites from *Pseudomonas aeruginosa* (Trunk et al., 2010). The deduced IUPAC consensus (B), regular expression (C), frequency matrix (D), position weight matrix (E) and sequence logo (F) are shown.

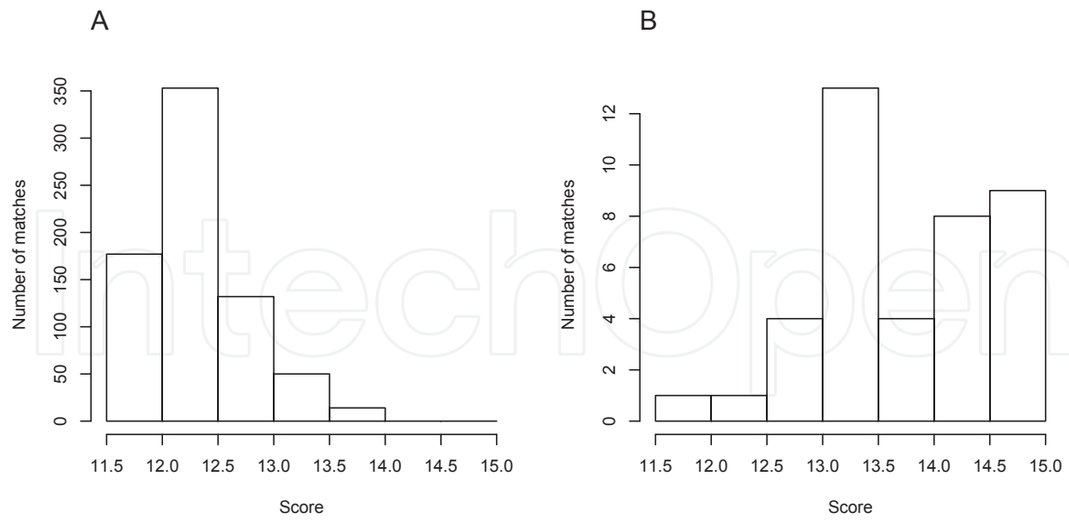


Fig. 2. Score distributions of false-positive matches (A) and true-positive matches (B) from a genome wide search in *P. aeruginosa* using the Anr PWM.

of four different classes (Fig. 3) which is called a two-by-two confusion matrix or contingency table (Fawcett, 2004).

	Dataset	
	Positive	Negative
Match	True-Positive	False-Positive
No Match	False-Negative	True-Negative

Fig. 3. A two-by-two confusion matrix illustrates all four possible outcomes of matches in the positive and in the negative dataset.

Thus, setting a cut-off score can be considered as important decision-making process. Instead of setting an arbitrary cut-off value it is possible to determine an optimized threshold. For that purpose, a number of statistical performance measurements for binary classification are available. Sensitivity Sn (or true-positive rate) measures the proportion of positive matches which are correctly identified at a given cut-off score c. Hereby, the positive matches include both the number of true-positives TP and false-negatives FN.

$$Sn(c) = \frac{TP}{TP + FN} \tag{4}$$

Similarly, specificity Sp (or true-negative rate) measures the proportion of correctly identified negative matches at a given cut-off score c where the amount of negative matches is the sum of true-negatives TN and false-positive FP.

$$Sp(c) = \frac{TN}{TN + FP} \tag{5}$$

This definition involves that the sensitivity and specificity plots as a function of the cut-off show opposite behaviour which results in an increase of specificity (get less false-positives) at the cost of sensitivity (find less true-positives) and vice versa (Fig. 4 A). A receiver operating characteristics (ROC) curve summarizes the classification performance in a plot of sensitivity versus (1-specificity). ROC curves are fundamental tools for the evaluation of the classification models. An optimal ROC curve would cross the upper left corner or coordinate (0,1) representing 100% sensitivity and specificity whereas a random guess would produce a point along the diagonal line (Fig. 4 A). Thus, the diagonal line divides the ROC space: points above the digonal represent good classification results, points below the line indicate poor results (Fawcett, 2004).

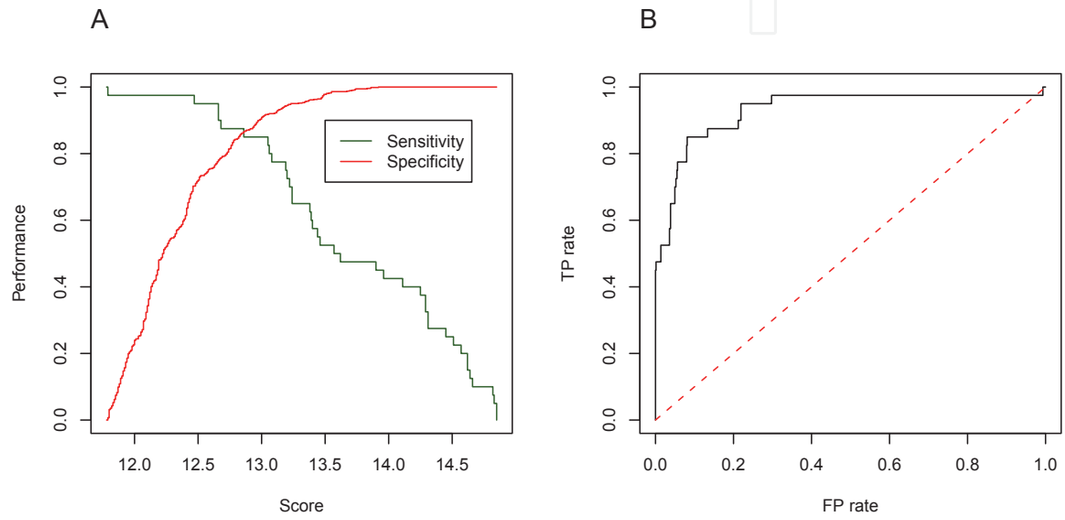


Fig. 4. Performance measurements for the prediction of the Anr regulon in *Pseudomonas aeruginosa*. (A) Sensitivity (green) and specificity (red) plot. (B) ROC graph.

An alternative way to optimize the performance of pattern matching and to produce stastistically significant results is the calculation of a *p*-value. A *p*-value depicts the likelihood to find a score that is as least as good by change. *P*-values can be either determined by simulation or estimated via a compound importance sampling approach (Oberto, 2010).
Finally, appropriate thresholds for pattern searches are determined as a tradeoff between sensitivity and specificity to maximize both values. Despite optimized cut-off values this approach can results in a poor sensitivity and a loss of 40-60% of known functional sites (Benítez-Bellón et al., 2002). In addition, the fact that false-predictions commonly exeeds true-predictions by several orders of magnitude (Fig. 2 B) was called **'futility theorem'** (Wasserman & Sandelin, 2004). Fortunately, there are many sophisticated approaches to overcome this problem in a reasonable way (see section 4).

4. Improvements to increase the accuracy of TFBS predictions

4.1 Modifications of the score

In several studies the information score was modified in different ways. One of the most critical points of equation 2 is that it postulates an equal nucleotide distribtuion of the target genome which is the case e.g. for *Escherichia coli* with a GC content of 51.8%. For this reason,

the calculation of the information content of motifs in genomes with highly biased nucleotide composition is likely to be over- or underestimated. A more generalized form that considers the background frequencies P_b is given in equation 6.

$$R(l) = -\sum_{b=A}^{T} f(b,l) \log_2 \frac{f(b,l)}{P_b} \tag{6}$$

This new term turned out to be the relative entropy or Kullback-Leibler distance (Stormo, 2000). An other promising approach deals with biased genome as a discrete channel of noise to discriminate a motif from its background (Schreiber & Brown, 2002). However, it was recently demonstrated, that the unmodified information score performs on average better than other alternatives (Erill & O'Neill, 2009). One reason might be, that binding sites shift towards the genome skew in a co-evolutionary process between TFs and its corresponding TFBSs.
Other modifications concern the way the score is computationally calculated. Since the information vector usually peeks at certain well conserved positions it is possible to get overestimated matches by forming the overall sum. For that purpose, it is useful to define a core region consisting of the highly conserved positions. Using this approach it is possible to realize the computation of the score in two steps. Potential matches have to pass first the core cut-off before they are evaluated by the overall cut-off score (Münch et al., 2005; Quandt et al., 1995).
Finally, it is possible to enhance the accuracy by combining multiple (independent) criterions. Apart from the pure sequence information, DNA exhibits distinct structural properties caused by interactions from neighboring nucleotides. This includes for example DNA curvature, flexibility and stability, amongst others. Structural DNA features are available as di- and trinucleotide scale values assigning a particular value to each possible nucleotide combination (Baldi & BaisnÃ©e, 2000). These values are derived from empirical measurements or theoretical approaches. The calculation of structural features within a DNA sequence stretch is usually performed by summing up and averaging the corresponding di- or trinucleotide scales. Prokaryotic promoters usually exhibit distinct structural features which imply that these DNA sequences are more curved and less flexible in comparision to coding regions. This feature is necessary in order to enable the melting of the DNA strands for the onset of transcription. In most bacterial promoters structural peaks are present around the position -40 upstream of the transcriptional start point (Pedersen et al., 2000). Structural features can provide distinct scores independent from PWM based sequence similarity scores. Recently, pattern matching was combined with a binding site model that was trained using 12 different structural properties (Meysman et al., 2011). In this approach, based on conditional random fields, it was shown, that the classification of matches was significantly improved. In a similar way, structural and chemical features of DNA decreased the number of false-positives in a supervised learning approach (Bauer et al., 2010).

4.2 Positional preference of TFBSs

Prokaryotic genomes usually consist of 6-14% non-coding DNA (Rogozin et al., 2002). In contrast to eukaryotes, the evolvement of non-coding regions appears to be determined primarily by the selective pressure to minimize the amount of non-functional DNA, while maintaining the essential TFBSs. Additionally, it was demonstrated in *Escherichia coli*, that many PWMs show a strong preference for matches in non-coding regions (Robison et al., 1998). Figure 5 A shows the distance of 1741 genomic TFBSs relative to the translational start site of the target gene. Only 3.6% of all TFBSs are located after the start codon within

the coding region. However, the largest amount of TFBSs is accumulated directly upstream. This is also demonstrated in the cumulative percentage of TFBSs against the distance to the translational start (Fig. 5 B). According to this result, a total of 75.3% and 87.9% of all TFBSs are located 200bp and 300bp upstream, respectively. Thus, prokaryotic promoters are usually short and it is reasonable to constrain searches to non-coding regions with a limit of a few hundred bp upstream to the translational start.

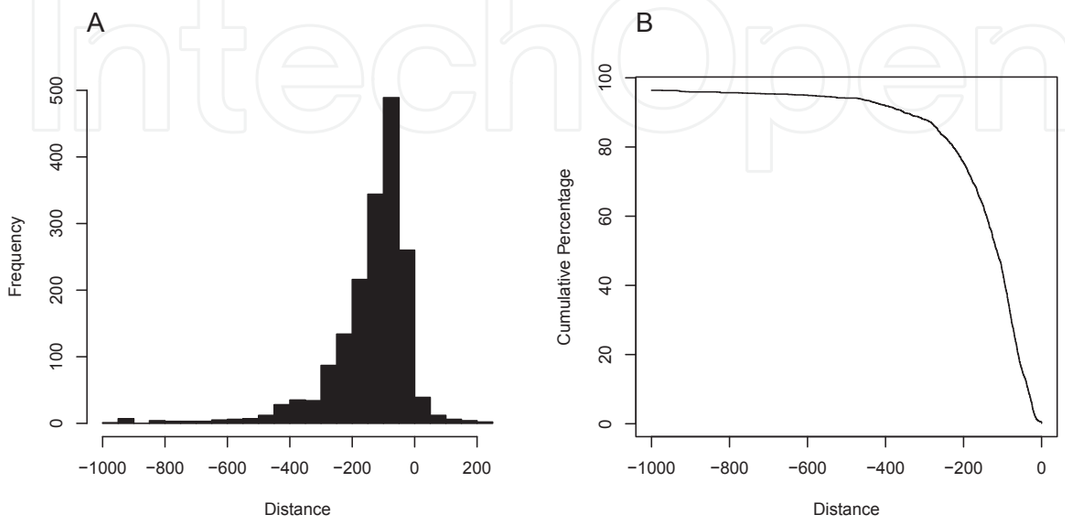


Fig. 5. Histogram of TFBS distances to the translational start site. The used dataset consisted of 1741 genomic TFBSs from various bacterial species taken from the PRODORIC database

4.3 Phylogenetic conservation of regulatory interactions

The large number of sequenced bacterial genomes offers comparative genomics approaches to predict and to analyze regulatory interactions. Similar to phylogenetic footprinting, highly conserved matches in promoter regions of paralogous genes are more likely to be functional targets than non-conserved matches (McCue et al., 2001). This is particulary important for the interspecies transfer of gene regulatory networks (Babu et al., 2006; Baumbach, 2010) but also for the scanning of new regulon members (Pérez et al., 2007). The utilization of pattern matching methods in combination with phylogenetic conservation is also called regulog analysis (Alkema et al., 2004). During a regulog analysis the relativ conservation score RCS is defined by the fraction of orthologs, that share the same potential TFBS.

$$RCS = \frac{orthologs_{observed}}{orthologs_{expected}} \tag{7}$$

In the first step of this and related approaches, the orthologous regulators and the corresponding target gene set are determined. This is often realized by bi-directional best BLAST hits (BBH) (Mushegian & Koonin, 1996). In the second step, conserved TFBSs are extracted via pattern matching or pattern discovery approaches. Predicted TFBSs with phylogenetic conservation can also be used to extend or to build new PWMs. Huge datasets based on phylogenetic reconstruction were generated in various groups of bacteria (Baumbach et al., 2009; Novichkov et al., 2010; Pérez et al., 2007). Further investigetion of regulon evolution revealed the availability of a core set of genes that is widely conserved

across related species and a variable set of target genes reflecting the degree of specialization (Browne et al., 2010; Dufour et al., 2010). However, it was shown, that the outlined approach is commonly only feasible between closely related clades which is due to the fact that TFs evolve rapidly and independently of their target genes (Babu et al., 2006). Morover, orthologous TFs in bacteria often have different functions and regulate different sets of genes (Price et al., 2007). In summary, a high *RCS* value for a TFBS match represents an independent score for the validation for a real functional targets while a low *RCS* does not necessarily rule out false-positive matches. The phylogenetic conservation approach represents a powerful approach to predict gene regulatory networks in highly related organisms and to get insights into the evolution of regulons.

5. Conclusion and outlook

In summary the genome-wide recognition of DNA patterns by computational methods is still a challanging task. However, major improvements in this field allow for reliable predictions in many cases. Especially the rising number of sequenced bacterial genomes in combination with data from high-throughput technologies offers many posibilities for the development of more sophisticated methods in comparative genomics approaches. Nevertheless, computational methods for TFBSs prediction can not replace wet-lab experiments but they can help to find new hypotheses that can be verified in an iterative process.

6. References

Alkema, W. B. L., Lenhard, B. & Wasserman, W. W. (2004). Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*., *Genome Res.* 14(7): 1362–1373.
URL: *http://dx.doi.org/10.1101/gr.2242604*

Babu, M. M., Teichmann, S. A. & Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks., *J Mol Biol* 358(2): 614–633.
URL: *http://dx.doi.org/10.1016/j.jmb.2006.02.019*

Baldi, P. & BaisnÃ©e, P. F. (2000). Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths., *Bioinformatics* 16(10): 865–889.

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* 6(11): e1001007.
URL: *http://dx.doi.org/10.1371/journal.pcbi.1001007*

Baumbach, J. (2010). On the power and limits of evolutionary conservation–unraveling bacterial gene regulatory networks., *Nucleic Acids Res* .
URL: *http://dx.doi.org/10.1093/nar/gkq699*

Baumbach, J., Wittkop, T., Kleindt, C. K. & Tauch, A. (2009). Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using coryneregnet., *Nat Protoc* 4(6): 992–1005.
URL: *http://dx.doi.org/10.1038/nprot.2009.81*

Benos, P. V., Bulyk, M. L. & Stormo, G. D. (2002). Additivity in protein-DNA interactions: how good an approximation is it?, *Nucleic Acids Res* 30(20): 4442–4451.

Benítez-Bellón, E., Moreno-Hagelsieb, G. & Collado-Vides, J. (2002). Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA., *Genome Biol* 3(3): 13.

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters., *J Mol Biol* 193(4): 723–750.

Betel, D. & Hogue, C. W. V. (2002). Kangaroo–a pattern-matching program for biological sequences., *BMC Bioinformatics* 3(1): 20.

Browne, P., Barret, M., O'Gara, F. & Morrissey, J. P. (2010). Computational prediction of the crc regulon identifies genus-wide and species-specific targets of catabolite repression control in *Pseudomonas* bacteria., *BMC Microbiol* 10: 300.
URL: *http://dx.doi.org/10.1186/1471-2180-10-300*

Day, W. H. & McMorris, F. R. (1992). Critical comparison of consensus methods for molecular sequences., *Nucleic Acids Res* 20(5): 1093–1099.

Dufour, Y. S., Kiley, P. J. & Donohue, T. J. (2010). Reconstruction of the core and extended regulons of global transcription factors., *PLoS Genet* 6(7): e1001027.
URL: *http://dx.doi.org/10.1371/journal.pgen.1001027*

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological sequence analysis*, Cambridge University Press.

Erill, I. & O'Neill, M. C. (2009). A reexamination of information theory-based methods for dna-binding site identification., *BMC Bioinformatics* 10: 57.
URL: *http://dx.doi.org/10.1186/1471-2105-10-57*

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers, *Technical report*, HP Laboratories.
URL: *http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf*

Fickett, J. W. (1996). Quantitative discrimination of MEF2 sites., *Mol Cell Biol* 16(1): 437–441.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J. S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, 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 Res* 39(Database issue): D98–105.
URL: *http://dx.doi.org/10.1093/nar/gkq1110*

Grote, A., Klein, J., Retter, I., Haddad, I., Behling, S., Bunk, B., Biegler, I., Yarmolinetz, S., Jahn, D. & Münch, R. (2009). PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes., *Nucleic Acids Res* 37(Database issue): D61–D65.
URL: *http://dx.doi.org/10.1093/nar/gkn837*

Hershberg, R., Bejerano, G., Santos-Zavaleta, A. & Margalit, H. (2001). PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites., *Nucleic Acids Res* 29(1): 277.

Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S. & Dubchak, I. (2007). RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes., *Nucleic Acids Res* 35(Database issue): D407–D412.
URL: *http://dx.doi.org/10.1093/nar/gkl865*

Loo, P. V. & Marynen, P. (2009). Computational methods for the detection of cis-regulatory modules., *Brief Bioinform* 10(5): 509–524.
URL: *http://dx.doi.org/10.1093/bib/bbp025*

McCue, L., Thompson, W., Carmack, C., Ryan, M. P., Liu, J. S., Derbyshire, V. & Lawrence, C. E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes., *Nucleic Acids Res* 29(3): 774–782.

Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J. & van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs., *Nucleic Acids Res* 39(3): 808–824.
URL: *http://dx.doi.org/10.1093/nar/gkq710*

Meysman, P., Dang, T. H., Laukens, K., Smet, R. D., 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* 39(2): e6.
URL: *http://dx.doi.org/10.1093/nar/gkq1071*

Münch, 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* 21(22): 4187–4189.
URL: *http://dx.doi.org/10.1093/bioinformatics/bti635*

Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes., *Proc Natl Acad Sci U S A* 93(19): 10268–10273.

NC-IUB (1985). Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984., *Eur J Biochem* 150(1): 1–5.

Novichkov, P. S., Laikova, O. N., Novichkova, E. S., Gelfand, M. S., Arkin, A. P., Dubchak, I. & Rodionov, D. A. (2010). RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes., *Nucleic Acids Res* 38(Database issue): D111–D118.
URL: *http://dx.doi.org/10.1093/nar/gkp894*

Oberto, J. (2010). Fitbar: a web tool for the robust prediction of prokaryotic regulons., *BMC Bioinformatics* 11: 554.
URL: *http://dx.doi.org/10.1186/1471-2105-11-554*

Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000). A DNA structural atlas for *Escherichia coli.*, *J Mol Biol* 299(4): 907–930.
URL: *http://dx.doi.org/10.1006/jmbi.2000.3787*

Pérez, A. G., Angarica, V. E., Vasconcelos, A. T. R. & Collado-Vides, J. (2007). Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes., *Nucleic Acids Res* 35(Database issue): D132–D136.
URL: *http://dx.doi.org/10.1093/nar/gkl800*

Price, M., Dehal, P. & Arkin, A. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli.*, *Genome Biol* 9(1): R4.
URL: *http://dx.doi.org/10.1186/gb-2008-9-1-r4*

Price, M. N., Dehal, P. S. & Arkin, A. P. (2007). Orthologous transcription factors in bacteria have different functions and regulate different genes., *PLoS Comput Biol* 3(9): 1739–1750.
URL: *http://dx.doi.org/10.1371/journal.pcbi.0030175*

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data., *Nucleic Acids Res* 23(23): 4878–4884.

Robison, K., McGuire, A. M. & 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(2): 241–254.

Rodionov, D. A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria., *Chem Rev* 107(8): 3467–3497.
URL: *http://dx.doi.org/10.1021/cr068309+*

Rogozin, I. B., Makarova, K. S., Natale, D. A., Spiridonov, A. N., Tatusov, R. L., Wolf, Y. I., Yin, J. & Koonin, E. V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes., *Nucleic Acids Res* 30(19): 4264–4271.

Sandve, G. K., Abul, O., Walseng, V. & DrablÃ¸s, F. (2007). Improved benchmarks for computational motif discovery., *BMC Bioinformatics* 8: 193.
URL: *http://dx.doi.org/10.1186/1471-2105-8-193*

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences., *Nucleic Acids Res* 18(20): 6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences., *J Mol Biol* 188(3): 415–431.

Schreiber, M. & Brown, C. (2002). Compensation for nucleotide bias in a genome by representation as a discrete channel with noise., *Bioinformatics* 18(4): 507–512.

Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. (2008). Dbtbs: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information., *Nucleic Acids Res* 36(Database issue): D93–D96.
URL: *http://dx.doi.org/10.1093/nar/gkm910*

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences., *Nucleic Acids Res* 12(1 Pt 2): 505–519.

Stormo, G. D. (2000). DNA binding sites: representation and discovery., *Bioinformatics* 16(1): 16–23.

Su, J., Teichmann, S. A. & Down, T. A. (2010). Assessing computational methods of cis-regulatory module prediction., *PLoS Comput Biol* 6(12): e1001020.
URL: *http://dx.doi.org/10.1371/journal.pcbi.1001020*

Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., RÃ©gnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites., *Nat Biotechnol* 23(1): 137–144.
URL: *http://dx.doi.org/10.1038/nbt1053*

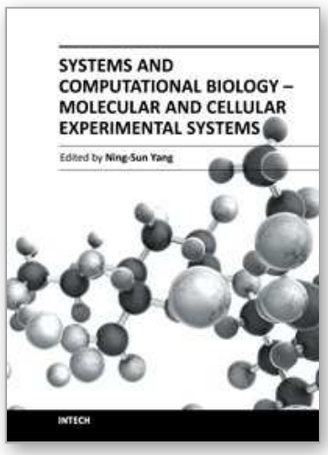
Trunk, K., Benkert, B., Quäck, N., Münch, R., Scheer, M., Garbe, J., Jänsch, L., Trost, M., Wehland, J., Buer, J., Jahn, M., Schobert, M. & Jahn, D. (2010). Anaerobic adaptation in *Pseudomonas aeruginosa*: definition of the Anr and Dnr regulons., *Environ Microbiol* 12(6): 1719–1733.
URL: *http://dx.doi.org/10.1111/j.1462-2920.2010.02252.x*

van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. (2009). Mechanisms and evolution of control logic in prokaryotic transcriptional regulation., *Microbiol Mol Biol Rev* 73(3): 481–509, Table of Contents.
URL: *http://dx.doi.org/10.1128/MMBR.00037-08*

Wasserman, W. W. & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements., *Nat Rev Genet* 5(4): 276–287.
URL: *http://dx.doi.org/10.1038/nrg1315*

Zhang, S., Xu, M., Li, S. & Su, Z. (2009). Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes., *Nucleic Acids Res* 37(10): e72.
URL: *http://dx.doi.org/10.1093/nar/gkp248*

Zhou, D. & Yang, R. (2006). Global analysis of gene transcription regulation in prokaryotes., *Cell Mol Life Sci* 63(19-20): 2260–2290.
URL: *http://dx.doi.org/10.1007/s00018-006-6184-6*



Systems and Computational Biology - Molecular and Cellular Experimental Systems

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
Publisher InTech
Published online 15, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

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

Richard Münch, Johannes Klein and Dieter Jahn (2011). Prediction and Analysis of Gene Regulatory Networks in Prokaryotic Genomes, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/prediction-and-analysis-of-gene-regulatory-networks-in-prokaryotic-genomes

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

IntechOpen