BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 22 2005, pages 4187–4189
doi:10.1093/bioinformatics/bti635

Genome analysis

Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes

Richard Münch[1], Karsten Hiller[1], Andreas Grote[1,2], Maurice Scheer[1,3], Johannes Klein[1], Max Schobert[1] and Dieter Jahn[1,*]

[1]Institute of Microbiology, Technical University Braunschweig, Spielmannstrasse 7, D-38106 Braunschweig, Germany, [2]Institute of Biochemical Engineering, Technical University Braunschweig, Gaußstrasse 17, 38106 Braunschweig, Germany and [3]Department of Informatics, University of Applied Sciences Wolfenbüttel, Am Exer 2, 38302 Wolfenbüttel, Germany

Received on May 11, 2005; revised and accepted on August 15, 2005
Advance Access publication August 18, 2005

ABSTRACT

Summary: A new online framework for the accurate and integrative prediction of transcription factor binding sites (TFBSs) in prokaryotes was developed. The system consists of three interconnected modules: (1) The PRODORIC database as a comprehensive data source and extensive collection of TFBSs with corresponding position weight matrices. (2) The pattern matching tool Virtual Footprint for the prediction of genome based regulons and for the analysis of individual promoter regions. (3) The interactive genome browser GBPro for the visualization of TFBS search results in their genomic context and links to gene and regulator-specific information in PRODORIC. The aim of this service is to provide researchers a free and easy to use collection of interconnected tools in the field of molecular microbiology, infection and systems biology.

Availability: http://www.prodoric.de/vfp

Contact: d.jahn@tu-bs.de

Table 1. Statistics of the PRODORIC database content with respect to TFBSs, regulated genes, regulons, PWMs and promoters

Organism	TFBSs	Genes	Regulons	PWMs	Promoters
Bacillus subtilis	662	662	77	65 (47)	410
Escherichia coli	1608	1015	81	84 (73)	719
Helicobacter pylori	9	11	2	2	9
Listeria monocytogenes	11	12	3	2	12
Pseudomonas aeruginosa	176	241	32	21 (18)	144
Streptococcus pyogenes	13	8	3	3 (2)	6
Others	38	45	19	5	36
Sum	2517	1994	217	182 (149)	1336

Status in August 2005.
In the case of several generated PWMs for one regulator, the non-redundant number is given in parentheses.

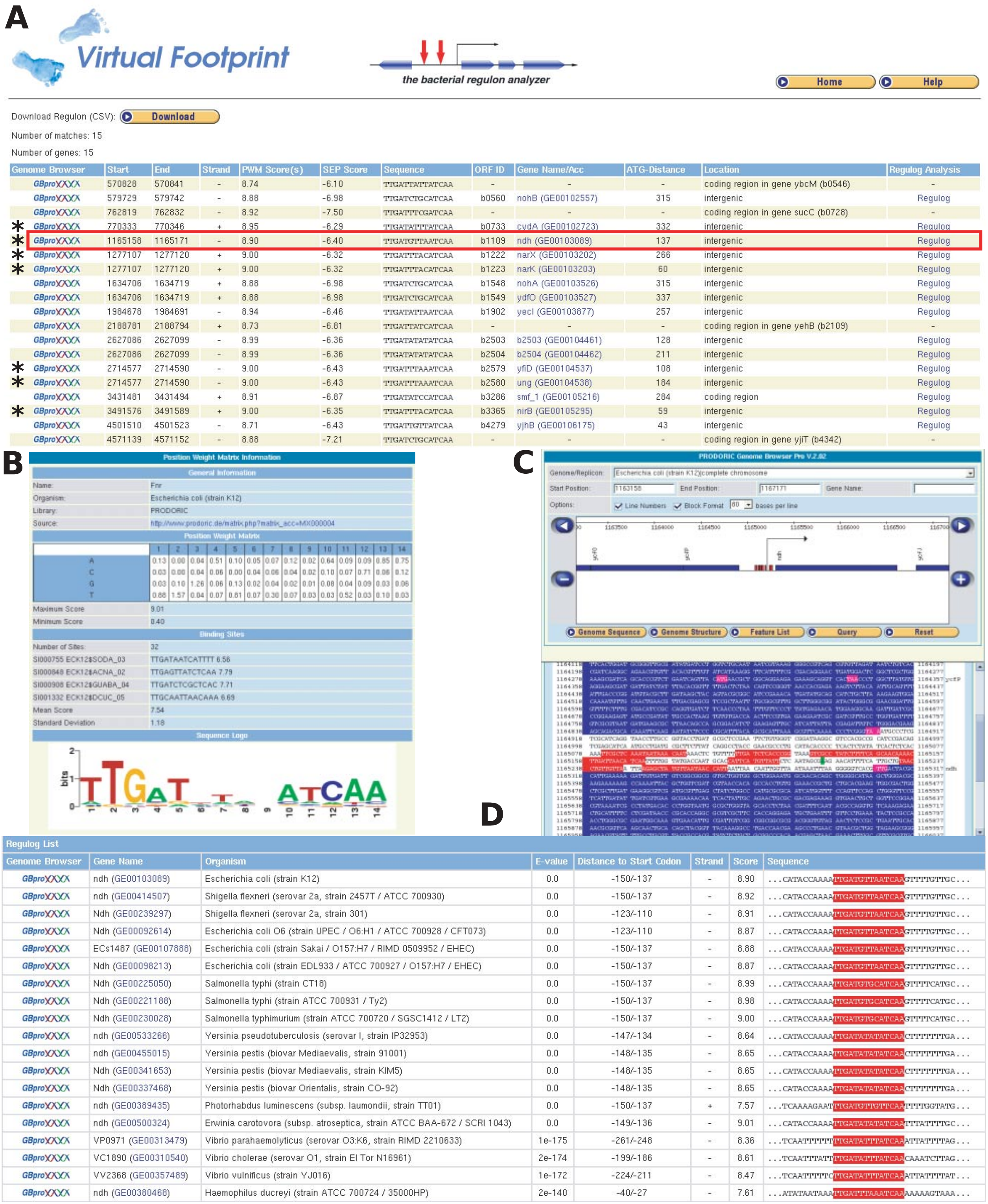
INTRODUCTION

The accurate prediction of transcription factor binding sites (TFBSs) and whole regulons is still a crucial step towards the understanding of complex regulatory networks in systems biology. Current bioinformatic methods of pattern recognition usually suffer from their low specificity. This often results in an accumulation of false-positive matches (Frech *et al*., 1997; Benitez-Bellon *et al*., 2002). Therefore, we developed a new framework for the straightforward evaluation and visualization of *in silico* results with focus on bacterial gene regulation. The user can interactively identify putative TFBSs using genome wide searches and immediately obtain detailed information on the promoters, corresponding genes, operons and encoded proteins found. This includes the genomic localization and detailed information about potentially regulated genes via links to the prokaryotic database of gene regulation (PRODORIC) database as well as relevant links to external sources. Moreover, it is possible to evaluate the matches obtained according to their phylogenetic conservation. The software is organized in three major interconnected components which are the PRODORIC database, the pattern search tool Virtual Footprint and the genome browser GBpro.

The PRODORIC database

The PRODORIC database is a comprehensive source of prokaryotic genomes and their underlying gene regulatory networks (Münch *et al* 2003). Among many other features it contains a compilation of over 2500 TFBSs from several bacterial species including their interacting transcriptional regulators. The data of PRODORIC are all based on experimental evidence which were manually extracted from the original literature. Using this huge collection of TFBSs over 170 species-specific position weight matrices (PWMs) were generated which serve as a library for the pattern matching tool Virtual Footprint. Besides the exclusive TFBS information, PRODORIC contains data of genes and proteins, promoter details, operon structures and links to relevant databases. Table 1 summarizes the data content of PRODORIC in the field of bacterial gene regulation. More recently the database was supplemented with additional information such as expression data from trans criptomics and proteomics experiments and metabolic

[*]To whom correspondence should be addressed.



Download Regulon (CSV): Download

Number of matches: 15

Number of genes: 15

Genome Browser	Start	End	Strand	PWM Score(s)	SEP Score	Sequence	ORF ID	Gene Name/Acc	ATG-Distance	Location	Regulog Analysis
GBpro	570828	570841	-	8.74	-6.10	TTGATTATTATCAA	-	-	-	coding region in gene ybcM (b0546)	-
GBpro	579729	579742	-	8.88	-6.98	TTGATCTGCATCAA	b0560	nohB (GE00102557)	315	intergenic	Regulog
GBpro	762819	762832	-	8.92	-7.50	TTGATTTCGATCAA	-	-	-	coding region in gene sucC (b0728)	-
* GBpro	770333	770346	+	8.95	-6.29	TTGATATTTATCAA	b0733	cydA (GE00102723)	332	intergenic	Regulog
* GBpro	1165158	1165171	-	8.90	-6.40	TTGATGTTAATCAA	b1109	ndh (GE00103089)	137	intergenic	Regulog
* GBpro	1277107	1277120	+	9.00	-6.32	TTGATTTACATCAA	b1222	narX (GE00103202)	266	intergenic	Regulog
* GBpro	1277107	1277120	+	9.00	-6.32	TTGATTTACATCAA	b1223	narK (GE00103203)	60	intergenic	Regulog
GBpro	1634706	1634719	+	8.88	-6.98	TTGATCTGCATCAA	b1548	nohA (GE00103526)	315	intergenic	Regulog
GBpro	1634706	1634719	+	8.88	-6.98	TTGATCTGCATCAA	b1549	ydfO (GE00103527)	337	intergenic	Regulog
GBpro	1984678	1984691	-	8.94	-6.46	TTGATATTAATCAA	b1902	yecI (GE00103877)	257	intergenic	Regulog
GBpro	2188781	2188794	+	8.73	-6.81	TTGATTATCATCAA	-	-	-	coding region in gene yehB (b2109)	-
GBpro	2627086	2627099	-	8.99	-6.36	TTGATATATATCAA	b2503	b2503 (GE00104461)	128	intergenic	Regulog
GBpro	2627086	2627099	-	8.99	-6.36	TTGATATATATCAA	b2504	b2504 (GE00104462)	211	intergenic	Regulog
* GBpro	2714577	2714590	-	9.00	-6.43	TTGATTTAAATCAA	b2579	yfiD (GE00104537)	108	intergenic	Regulog
* GBpro	2714577	2714590	-	9.00	-6.43	TTGATTTAAATCAA	b2580	ung (GE00104538)	184	intergenic	Regulog
GBpro	3431481	3431494	+	8.91	-6.87	TTGATATCCATCAA	b3286	smf_1 (GE00105216)	284	coding region	Regulog
* GBpro	3491576	3491589	+	9.00	-6.35	TTGATTTACATCAA	b3365	nirB (GE00105295)	59	intergenic	Regulog
GBpro	4501510	4501523	-	8.71	-6.43	TTGATTGTTATCAA	b4279	yjhB (GE00106175)	43	intergenic	Regulog
GBpro	4571139	4571152	-	8.88	-7.21	TTGATCTGCATCAA	-	-	-	coding region in gene yjiT (b4342)	-

Position Weight Matrix Information

General Information

Name:	Fnr
Organism:	Escherichia coli (strain K12)
Library:	PRODORIC
Source:	http://www.prodoric.de/matrix.php?matrix_acc=MX000004

Position Weight Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0.13	0.00	0.04	0.51	0.10	0.05	0.07	0.12	0.02	0.64	0.09	0.09	0.85	0.75
C	0.03	0.00	0.04	0.06	0.00	0.04	0.06	0.04	0.02	0.10	0.07	0.71	0.06	0.12
G	0.03	0.10	1.26	0.06	0.13	0.02	0.04	0.02	0.01	0.08	0.04	0.09	0.03	0.06
T	0.88	1.57	0.04	0.07	0.81	0.07	0.30	0.07	0.03	0.03	0.52	0.03	0.10	0.03

Maximum Score	9.01
Minimum Score	0.40

Binding Sites

Number of Sites:	32
SI000755 ECK12$SODA_03	TTGATAATCATTTT 6.56
SI000848 ECK12$ACNA_02	TTGAGTTATCTCAA 7.79
SI000908 ECK12$GUABA_04	TTGATCTCGCTCAC 7.71
SI001332 ECK12$DCUC_05	TTGCAATTAACAAA 6.69
Mean Score	7.54
Standard Deviation	1.18

Genome/Replicon:	Escherichia coli (strain K12)(complete chromosome		
Start Position:	1163158	End Position:	1167171
Gene Name:			
Options:	☑ Line Numbers ☑ Block Format 60 bases per line		

Regulog List

Genome Browser	Gene Name	Organism	E-value	Distance to Start Codon	Strand	Score	Sequence
GBpro	ndh (GE00103089)	Escherichia coli (strain K12)	0.0	-150/-137	-	8.90	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	ndh (GE00414507)	Shigella flexneri (serovar 2a, strain 2457T / ATCC 700930)	0.0	-150/-137	-	8.92	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	Ndh (GE00239297)	Shigella flexneri (serovar 2a, strain 301)	0.0	-123/-110	-	8.91	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	Ndh (GE00092614)	Escherichia coli O6 (strain UPEC / O6:H1 / ATCC 700928 / CFT073)	0.0	-123/-110	-	8.87	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	ECs1487 (GE00107888)	Escherichia coli (strain Sakai / O157:H7 / RIMD 0509952 / EHEC)	0.0	-150/-137	-	8.88	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	Ndh (GE00098213)	Escherichia coli (strain EDL933 / ATCC 700927 / O157:H7 / EHEC)	0.0	-150/-137	-	8.87	...CATACCAAAATTGATGTTAATCAAGTTTTGTTGC...
GBpro	Ndh (GE00225050)	Salmonella typhi (strain CT18)	0.0	-150/-137	-	8.99	...CATACCAAAATTGATGTGCATCAAGTTTTCATGC...
GBpro	Ndh (GE00221188)	Salmonella typhi (strain ATCC 700931 / Ty2)	0.0	-150/-137	-	8.98	...CATACCAAAATTGATGTGCATCAAGTTTTCATGC...
GBpro	Ndh (GE00230028)	Salmonella typhimurium (strain ATCC 700720 / SGSC1412 / LT2)	0.0	-150/-137	-	9.00	...CATACCAAAATTGATGTGCATCAAGTTTTCATGC...
GBpro	ndh (GE00533266)	Yersinia pseudotuberculosis (serovar I, strain IP32953)	0.0	-147/-134	-	8.64	...CATACCAAAATTGATATATATCAACTTTTTTTGA...
GBpro	ndh (GE00455015)	Yersinia pestis (biovar Mediaevalis, strain 91001)	0.0	-148/-135	-	8.65	...CATACCAAAATTGATATATATCAACTTTTTTTGA...
GBpro	Ndh (GE00341653)	Yersinia pestis (biovar Mediaevalis, strain KIM5)	0.0	-148/-135	-	8.65	...CATACCAAAATTGATATATATCAACTTTTTTTGA...
GBpro	Ndh (GE00337468)	Yersinia pestis (biovar Orientalis, strain CO-92)	0.0	-148/-135	-	8.65	...CATACCAAAATTGATATATATCAACTTTTTTTGA...
GBpro	ndh (GE00389435)	Photorhabdus luminescens (subsp. laumondii, strain TT01)	0.0	-150/-137	+	7.57	...TCAAAAGAATTTGATGTTGTTCAATTTTGGTATG...
GBpro	ndh (GE00500324)	Erwinia carotovora (subsp. atroseptica, strain ATCC BAA-672 / SCRI 1043)	0.0	-149/-136	-	9.01	...CATACCAAAATTGATATATATCAATTTATTTTGC...
GBpro	VP0971 (GE00313479)	Vibrio parahaemolyticus (serovar O3:K6, strain RIMD 2210633)	1e-175	-261/-248	-	8.36	...TCAATTTTTTTTGATATTTATCAAATTATTTTAG...
GBpro	VC1890 (GE00310540)	Vibrio cholerae (serovar O1, strain El Tor N16961)	2e-174	-199/-186	-	8.61	...TCAATTTATTTTGATATTTATCAACAAATCTTAG...
GBpro	VV2368 (GE00357489)	Vibrio vulnificus (strain YJ016)	1e-172	-224/-211	-	8.47	...TCAATTTTTCTTGATATTTATCAAATTATTTTAT...
GBpro	ndh (GE00380468)	Haemophilus ducreyi (strain ATCC 700724 / 35000HP)	2e-140	-40/-27	-	7.61	...ATATAATAAATTGATTTAAATCAAAAAAAGTAAA...

Fig. 1. Example for the application of Virtual Footprint for a genome wide prediction of Fnr (fumerate nitrate reduction regulatory protein) binding sites in *Escherichia coli*. (**A**) List of high scoring matches including their genomic positions, scores, assigned downstream genes, with links to PRODORIC, the genome browser GBpro and regulog analyses. Matches marked with asterisks are documented in the PRODORIC database. One of the matches upstream of the *ndh* gene (respiratory NADH dehydrogenase) is framed (**B**) Fnr PWM with sequence logo (the number of used regulatory sequences was reduced due to space limitations). (**C**) Genome browser view of the genomic context of *ndh*. (**D**) Regulog analysis of the Fnr binding site upstream of *ndh*. The resulting list provides links to the genome browser and to the gene entries of PRODORIC.

networks. This integrated approach makes PRODORIC well suited as a platform for systems biology in prokaryotes.

Virtual Footprint—pattern matcher

The new pattern search tool Virtual Footprint offers fast searches of complex DNA patterns in whole bacterial genomes. Usually the search pattern is defined as PWM provided by PRODORIC (Fig. 1B). We also added PWMs from other resources (Robison *et al*., 1998; Salgado *et al*., 2004). However, in some cases, e.g. when sufficient sequence data are not available, it is necessary to use other pattern definitions (Stormo, 2000). Therefore, we implemented search algorithms using IUPAC consensi and regular expressions (Betel *et al*., 2002). Some TFBSs are not only variable in their sequence conservation but also in their sequence lengths. Common examples are the occurrence of two half-sites separated by a variable spacer, the conserved -10 and -35 hexamers found in σ^{70} regulated promoters and other so, called composite elements (Kel-Margoulis *et al*., 2000). Therefore, Virtual Footprint allows the definition of bipartite patterns via the combination of up to two subpatterns separated by a variable spacer. Different pattern types can be freely combined e.g. a PWM with a IUPAC string. This enables a flexible definition of search patterns. The list of obtained matches can be evaluated by several different *in silico* approaches in accordance with their genomic context. Usually matches are directly linked to the downstream genes. Identified genes or operons are linked to the PRODORIC database and the genome browser GBpro (Fig. 1). If it is not possible to assign a gene to a match, the corresponding genomic location is specified. For a search the size of the upstream region (distance to the start codon) can be defined, the pattern orientation can be selected and matches in coding regions can be excluded. Matches can be further evaluated by analyzing upstream regions of orthologous genes for the same pattern (Fig. 1D). In this case orthologous sequence stretches from different genomes are extracted by the use of BLAST (Altschul *et al*., 1990) and then analyzed via Virtual Footprint. This kind of investigation is also called regulog analysis (Alkema *et al*., 2004). Furthermore, the GC-content of a promoter region and the resulting stacking energy plot shown by the genome browser can help to evaluate matches since functional targets should be localized in chromosomal regions where the GC-content and stacking energy are expected to be below the average. In addition to the whole genome search, Virtual Footprint allows the analysis of single promoter regions. In this case the upstream sequence of a gene or a pasted user defined sequence is compared with all PWMs provided by PRODORIC. Virtual Footprint has been successfully applied to define the ResD regulon of *Bacillus subtilis* (Härtig *et al*., 2004) and to detect split tRNA genes in *Nanoarchaeum equitans* (Randau *et al*., 2005). The program offers many options and settings not described here. A detailed description is available in the online help of the program.

GBpro—Genome Browser

GBpro is a genome browser for an interactive navigation through all bacterial genomes available in PRODORIC. Genes, promoters and binding sites are displayed in parallel as graphical maps and highlighted sequences. Optionally, the GC-content and stacking energy of a DNA sequence of interest can be visualized. All results of Virtual Footprint are directly linked to GBpro and can thus be visualized in their genomic context (Fig. 1C). Similarly, genes and TFBSs present in PRODORIC are directly linked to this GBpro.

ACKNOWLEDGEMENTS

We would like to thank Denise Wätzlich and Lorenz Reimer for technical support and Dr Barbara Schulz for critical proofreading of the manuscript. We also thank Karin and Werner Müunch for assistance in web design. This work was funded by the German Federal Ministry of Education and Research (BMBF) for the Bioinformatics Competence Center 'Intergenomics' (Grant No. 031U110A/031U210A).

Conflict of Interest: none declared.

REFERENCES

Alkema,W.B.L. *et al*. (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. *Genome Res*., **14**, 1362–1373.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

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

Benitez-Bellon,E. *et al*. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA. *Genome Biol*., **3**, 13.

Betel,D. and Hogue,C. (2002) Kangaroo—A pattern-matching program for biological sequences. *BMC Bioinformatics*, **3**, 20.

Frech,K. *et al*. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci*., **22**, 103–104.

Härtig,E. *et al*. (2004) *Bacillus subtilis* ResD induces expression of the potential regulatory genes yclJK upon oxygen limitation. *J. Bacteriol*., **186**, 6477–6484.

Kel-Margoulis,O.V. *et al*. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res*., **28**, 311–315.

Münch,R. *et al*. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*., **31**, 266–269.

Randau,L. *et al*. (2005) *Nanoarchaeum equitans* creates functional tRNA from separate genes for their 5′- and 3′-halves. *Nature*, **433**, 537–541.

Robison,K. *et al*. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K12 genome. *J. Mol. Biol*., **284**, 241–254.

Salgado,H. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K12. *Nucleic Acids Res*., **32**, 303–306.

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