Available online at www.sciencedirect.com

ScienceDirect

Biochemical and Biophysical Research Communications 357 (2007) 861–864

BBRC

www.elsevier.com/locate/ybbrc

PREDetector: A new tool to identify regulatory elements in bacterial genomes

Samuel Hiard [a], Raphaël Marée [b], Séverine Colson [c], Paul A. Hoskisson [d], Fritz Titgemeyer [e], Gilles P. van Wezel [f], Bernard Joris [c], Louis Wehenkel [a], Sébastien Rigali [c,f,*]

[a] *Department of Electrical Engineering and Computer Science, University of Liège, B28 Systems and Modeling, Grande Traverse 10, 4000 Liège, Belgium*
[b] *GIGA: plateforme bioinformatique, University of Liège, B28 Systems and Modeling, Grande Traverse 10, 4000 Liège, Belgium*
[c] *Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6a, Sart-Tilman, B-4000 Liège, Belgium*
[d] *Department of Molecular and Cell Biology, Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK*
[e] *Lehrstuhl für Mikrobiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Staudtstrasse 5, 91058 Erlangen, Germany*
[f] *Microbial Development, Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300RA Leiden, The Netherlands*

Received 27 March 2007
Available online 12 April 2007

Abstract

In the post-genomic area, the prediction of transcription factor regulons by position weight matrix-based programmes is a powerful approach to decipher biological pathways and to modelize regulatory networks in bacteria. The main difficulty once a regulon prediction is available is to estimate its reliability prior to start expensive experimental validations and therefore trying to find a way how to identify true positive hits from an endless list of potential target genes of a regulatory protein. Here we introduce PREDetector (Prokaryotic Regulatory Elements Detector), a tool developed for predicting regulons of DNA-binding proteins in bacterial genomes that, beside the automatic prediction, scoring and positioning of potential binding sites and their respective target genes in annotated bacterial genomes, it also provides an easy way to estimate the thresholds where to find reliable possible new target genes. PREDetector can be downloaded freely at http://www.montefiore.ulg.ac.be/~hiard/PreDetector/PreDetector.php.

Keywords: Regulon prediction; Transcriptional regulation; Regulatory networks; Position weight matrix; DNA-binding motif

Genome sequences are a mine of information to estimate the natural predisposition of a microorganism to face and respond to particular ecological niches or can be regarded as valuable resources for interrogation to specific biotechnological ends. However, beyond these basic genetic data, the assessment of the real metabolic and physiological potentialities requires intensive investigations on how the living cell senses environmental signals and transmits messages to regulatory authorities that control genes expression. The characterisation of a regulon, i.e. the transcription factor(s) (TF), *cis*-acting element(s), ligand affecting the DNA-binding ability, the set of target genes and the controlled biological processes, is crucial to understand all living organisms. Deciphering the *cis–trans* relationships that weave a regulatory network is considered as the first step towards this aim. Once regulatory DNA sequences have been demonstrated as targets for a specific transcription factor, their conserved signature is generally described by position weight matrices (PWMs) which specify the frequency distribution of nucleotides at each position of the TF *cis*-acting elements. Several PWMs based web tools (such as Target Explorer [1], Virtual Footprint [2], or PredictRegulon [3]) have directed experimental investigation by highlighting potential new target genes of a transcription factor [4], defining a regulon in closely related bacterial species, and also have contributed to revealing

---

[*] Corresponding author. Address: Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6a, Sart-Tilman, B-4000 Liège, Belgium.

E-mail addresses: S.Hiard@ulg.ac.be (S. Hiard), L.Wehenkel@ulg.ac.be (L. Wehenkel), Srigali@ulg.ac.be (S. Rigali).

0006-291X/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.bbrc.2007.03.180

new *cis–trans* relationships [5]. The diversity and complexity of *in vivo cis–trans* relationships make computational predictions of transcription factors regulons hazardous. Two main difficulties are to identify true positive hits from an endless list of potential target genes and, at the opposite, some programmes are too restrictive and do not allow generating predictions beyond a certain reliability threshold. Often, it is user's experience and knowledge in the field that is the best, and probably, the unique way to categorize a predicted motif in the list of target hits that deserve further investigation for experimental validation. Because this specific scientific expertise is unfortunately almost impossible to integrate within software, it is therefore too restrictive to use computational programmes that do not leave the possibility to scientists to retrieve information at thresholds generating high levels of false positive hits. The priority and challenge of PREDetector (Prokaryotic Regulatory Elements Detector) was to offer a programme which, all at once, would provide an easy way to estimate the reliability of the predictions, and beyond the identification of strongly reliable *cis*-acting elements, would guarantee users the possibility to access information among the list of predicted sites with scores generally regarded too low to be reliable. In addition, with PREDetector users have also access to sites located within coding regions and regions in between two translation Stop codons and therefore generally omitted during investigations because systematically considered with no functional elements. The characteristics and the possibilities offered to users by PREDetector are explained and illustrated below with the prediction of the ScbR (γ-butyrolactone binding protein) regulon of *Streptomyces avermitilis* using experimentally validated ScbR binding sites of *Streptomyces coelicolor* [6,7].

Software description

Generation of weight matrices

PREDetector is structurally divided into two sections called "Weight Matrix Creation" and "Regulon Prediction", respectively. In the "Weight Matrix Creation" section (Supplementary Fig. 1), users can build a PWM from a set of DNA sequences. Sequences of equal (L) lengths are inserted in fasta format. Once the set of input sequences has been inserted, a consensus is deduced from the multiple alignment which is also converted into a weight matrix using the expression [8]:

$$\text{weight}_{i,j} = \ln\left\{\frac{[(n_{i,j}+p_i)/(N+1)]}{p_i}\right\} \sim \ln\left(\frac{f_{i,j}}{p_i}\right)$$

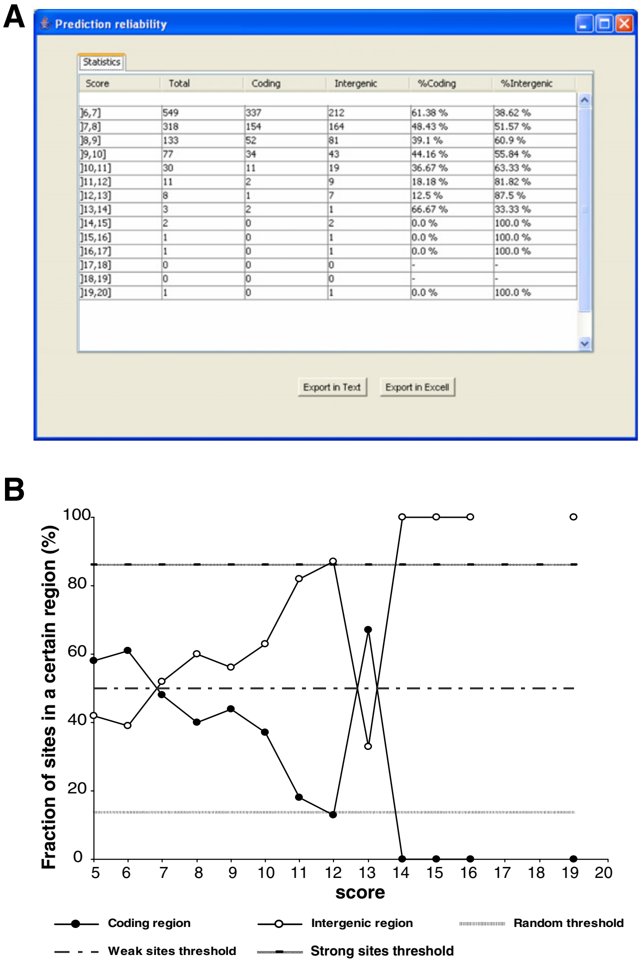
where N is the total number of sequences in the alignment, $n_{i,j}$ is the number of times nucleotide i is observed in position j of the alignment, $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j, and p_i is the *a priori*: the frequency of base i in the genome (for example, 0.25 for each nucleotide in a genome where the G + C content is 50%). For the best prediction accuracy, users are encouraged to fix themselves the *a priori* according to the nucleotide content of the genome where DNA-binding sites have been characterized. The generated matrices can be saved and used for future computational predictions.

Regulon prediction

The search for potential binding sites of the regulatory protein starts with the selection of one of the saved position weight matrices from the user's library and the definition of the cut-off score (Supplementary Fig. 2). The lowest score among the input sequences used to build a matrix is fixed by default as the recommended cut-off score for this matrix. Our web tool allows users to modify the cut-off score (see below estimation of the reliability of the predictions). PREDetector is able to scan either complete or selected regions of bacterial genomes available in the GenBank database [9]. Once the options have been set, PREDetector scans the selected genome sequences and classifies the predicted target DNA motifs according to their localisation in the genome. This includes coding sequences or intergenic sequences, which can be classified as (1) regulatory regions (where regulatory elements are predicted to be found), (2) upstream regions (any region upstream of a translational start codon), and (3) terminator regions (in PREDetector the terminator region terminology is only used to indicate regions between two translational stop codons) (Supplementary Fig. 3A). Predictions results are distributed among these four genome localization categories (Supplementary Fig. 3B–E).

Estimation of the reliability of the predictions

Besides the amount and the variability of the binding-sites available for the studied transcription factor, the efficiency of a prediction mostly depends on the choice of the threshold of the cut-off score. A cut-off score that is set too high could of course prevent the identification of truly functional elements, while a cut-off score that is too low will ensure a high false-positive rate. In PREDetector, the lowest score among the input sequences is considered by default as the cut-off score. Once the first round of prediction has been completed, our web tool allows to readily modify and check the appropriate cut-off score to use in any search. A methodology to set the threshold to predict strong and weak sites was described by Tan and collaborators [10]. In summary, given the observation that functional regulatory sites usually are located upstream transcription units (TUs), they considered that the threshold for the cut-off to predict new strong sites must have a score at which most sites are located upstream of TUs, and the threshold for weak sites must have a score at which approximately half of all sites are located upstream of TUs (Fig. 1A and B). At low cut-off scores, almost all sites are located within TUs, indicating a high false-positive rate. A simple approach to determine the inferior limit of weak sites is to take count of the amount of upstream regions in the consid-



Score	Total	Coding	Intergenic	%Coding	%Intergenic
]6,7]	549	337	212	61.38 %	38.62 %
]7,8]	318	154	164	48.43 %	51.57 %
]8,9]	133	52	81	39.1 %	60.9 %
]9,10]	77	34	43	44.16 %	55.84 %
]10,11]	30	11	19	36.67 %	63.33 %
]11,12]	11	2	9	18.18 %	81.82 %
]12,13]	8	1	7	12.5 %	87.5 %
]13,14]	3	2	1	66.67 %	33.33 %
]14,15]	2	0	2	0.0 %	100.0 %
]15,16]	1	0	1	0.0 %	100.0 %
]16,17]	1	0	1	0.0 %	100.0 %
]17,18]	0	0	0	-	-
]18,19]	0	0	0	-	-
]19,20]	1	0	1	0.0 %	100.0 %

Fig. 1. Estimation of the reliability of the predictions: determination of the appropriate cut-off score. The *cis*-acting sequences bound by ScbR of *Streptomyces coelicolor* [6,7] were used as training set to generate the *scbR* weight matrix. (A) Prediction reliability output: fractions of sites located either within intergenic or coding regions in *S. avermitilis*. (B) Graphic representation of the Prediction reliability output. Illustration of the determination of strong, weak, and random sites' thresholds according to Tan et al. [10] methodology. The amount of coding and intergenic regions in *S. avermitilis* is 86 and 14%, respectively. Therefore, the threshold to predict new strong sites is fixed at scores at which 86% of the predicted sites are located within intergenic regions. By convention, the threshold for weak sites is fixed at scores at which about the half of all sites are located within intergenic regions. At low cut-off scores, most of predicted sites are located within TUs and are closer to the threshold for a random localization. This threshold is around 14% in *S. avermitilis* which means that sites with random localization occur 14% of the time within intergenic regions.

ered organism. In *S. avermitilis* for instance, this amount is ~14% of the genome, which means that sites with random localization occur ~14% of the time upstream TUs. This threshold (in *S. avermitilis* ~14%) will be called hereafter the "random threshold" (Fig. 1B). In this case all sites predicted with scores at which ~86% or more of sites are within TUs should not be selected for the list of possible new target sites of a particular transcription factor. Thus from the "Prediction Reliability" output, PREDetector offers users the possibility to check the distribution of predicted sites, i.e. within intergenic or coding regions, according to their score (Fig. 1B). Therefore, users know which categories (strong, weak, or random) the predicted target sequences belong and can consequently adapt the cut-off score from each set of input sequences and each genome.

Conclusion

This paper describes a new tool—PREDetector—suitable for genome-wide prediction of potential *cis*-acting elements that aims to compile most of the features offered by several tools already available, avoiding the repetition of prediction searches and retrieving complementary data from different sources. PREDetector automatically (1) predicts, scores and positions potential binding sites and their respective target genes, (2) includes the downstream co-regulated genes, (3) extends the predictions to coding sequences and terminator regions, (4) saves users' private matrices and allows predictions in other genomes, and (5) provides an easy way to estimate the reliability of the predictions. PREDetector is continuously under development and suggestions for improvements are welcome (contact Samuel Hiard, S.Hiard@ulg.ac.be and Louis Wehenkel, L.Wehenkel@ulg.ac.be). PREDetector is available at http://www.montefiore.ulg.ac.be/~hiard/PreDetector/PreDetector.php [11].

Acknowledgments

This work presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. S.H. is supported by a grant from the patrimoine of the University of Liège. S.C. is a research fellow of the Fund of Research in Industry and Agriculture (FRIA) and S.R. was supported by the patrimoine of the University of Liège (Belgium) and an ECHO grant from the Netherlands Organization for Scientific Research (NWO).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.03.180.

References

[1] A. Sosinsky, C.P. Bonin, R.S. Mann, B. Honig, Target explorer: an automated tool for the identification of new target genes for a specified set of transcription factors, Nucleic Acids Res. 31 (2003) 3589–3592.
[2] R. Munch, K. Hiller, A. Grote, M. Scheer, J. Klein, M. Schobert, D. Jahn, Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes, Bioinformatics 21 (2005) 4187–4189.
[3] S. Yellaboina, J. Seshadri, M.S. Kumar, A. Ranjan, PredictRegulon: a web server for the prediction of the regulatory protein binding sites

and operons in prokaryote genomes, Nucleic Acids Res. 32 (2004) W318–W320.

[4] S. Rigali, M. Schlicht, P. Hoskisson, H. Nothaft, M. Merzbacher, B. Joris, F. Titgemeyer, Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new *cis*/*trans* relationships, Nucleic Acids Res. 32 (2004) 3418–3426.

[5] S. Rigali, H. Nothaft, E.E. Noens, M. Schlicht, S. Colson, M. Muller, B. Joris, H.K. Koerten, D.A. Hopwood, F. Titgemeyer, et al., The sugar phosphotransferase system of *Streptomyces coelicolor* is regulated by the GntR-family regulator DasR and links *N*-acetylglucosamine metabolism to the control of development, Mol. Microbiol. 61 (2006) 237–1251.

[6] E. Takano, R. Chakraburtty, T. Nihira, Y. Yamada, M.J. Bibb, A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in *Streptomyces coelicolor* A3(2), Mol. Microbiol. 41 (2001) 1015–1028.

[7] E. Takano, H. Kinoshita, V. Mersinias, G. Bucca, G. Hotchkiss, T. Nihira, C.P. Smith, M. Bibb, W. Wohlleben, K. Chater, A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of *Streptomyces coelicolor*, Mol. Microbiol. 56 (2005) 465–479.

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

[9] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, GenBank, Nucleic Acids Res. 34 (2006) D16–D20.

[10] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides, G.D. Stormo, A comparative genomics approach to prediction of new members of regulons, Genome Res. 11 (2001) 566–584.

[11] PreDetector http://www.montefiore.ulg.ac.be/~hiard/PreDetector/PreDetector.php.