Molecular BioSystems

OPINION

On the Necessity and Biological Significance of Threshold-free Regulon Prediction Outputs

Cite this: DOI: 10.1039/x0xx00000x

Received 13th August 2014,

DOI: 10.1039/x0xx00000x

www.rsc.org/

Sébastien Rigali[ac], Renaud Nivelle[ab], and Pierre Tocquin[b]

The *in silico* prediction of *cis*-acting elements in a genome is an efficient way to quickly obtain an overview of the biological processes controlled by a *trans*-acting factor, and connections between regulatory networks. Several regulon prediction web tools are available, designed to identify DNA motifs predicted to be bound by transcription factors using position weight matrix-based algorithms. In this paper we expose and discuss the conflicting objectives of software creators (bioinformaticians) and software users (biologists), who aim for reliable and exhaustive prediction outputs, respectively. Software makers, concerned with providing tools that minimise the number of false positive hits, often impose a stringent threshold score for a sequence to be included in the list of the putative *cis*-acting sites. This rigidity eventually results in the identification of strongly reliable but largely straightforward sites, i.e. those associated with genes already anticipated to be targeted by the studied transcription factor. Importantly, this biased identification of strongly bound sequences contrasts with the biological reality where, in many circumstances, a weak DNA-protein interaction is required for the appropriate gene's expression. We show here a series of transcriptionally controlled systems involving weakly bound *cis*-acting elements that could never have been discovered because of the policy of preventing software users from modifying the screening parameters. Proposing only trustworthy prediction outputs thus prevents biologists from fully utilising their knowledge background and deciding to analyse statistically irrelevant hits that could nonetheless be potentially involved in subtle, unexpected, though essential *cis*-*trans* relationships.

Regulon prediction web tools and threshold score apprehension

The physical interaction between regulatory DNA-binding proteins and their cognate DNA sequences directs the spatio-temporal and the elicitor-dependent expression of genes whose product is only required either at a certain moment or under specific environmental conditions. Bioinformatic programs designed to identify the *cis*-acting elements bound by a transcription factor (TF) have been demonstrated to be efficient tools in System Biology, able to quickly unveil genes whose expression is associated with specific or interconnected biological processes. As we have now undeniably entered an era of low-cost DNA-sequencing in which novel and fully annotated genomes are deposited in specialised databases on a daily basis, the *in silico* prediction of a TF regulon has become an examination 'reflex' preliminary to expensive *in vivo* and *in vitro* genome-wide investigations. The growing popularity of such computational work can be inferred from the number of software products used for predicting regulons, as well as by the regular updates of the latter since they first became publicly accessible.

Typically, in a supervised motif finding approach[1], the software user begins the regulon prediction process by creating a position weight matrix (PWM) that attempts to best represent the tolerated variability of a series of DNA sequences known to be bound by a TF[2]. With minor modifications depending on the algorithm used, PWMs are obtained by attributing a score to any nucleotide *i* at position *j* of an *L* length sequence, based on its frequency of appearance in the TF-binding sites that fed the algorithm. The score of an *L* length sequence is the sum of individual scores of each nucleotide composing the sequence. The PWM will then serve to scan the full or partial genome sequence in order to identify genes neighbouring identical or similar *L* length DNA sequences presumed to fall under the expression control of the studied TF. The number of sequences in the regulon prediction output list, and thus the ratio between false positive hits and truly bound sequences *in vivo*, will depend on the threshold or cut-off score fixed by the software user, where this is allowed by the software developer.

Software creators aim for reliable prediction outputs

From the standpoint of the software creator (for whom the output is the final result, with no further development), the legitimacy of a regulon prediction program depends on its ability to provide primarily positive hits. The guaranteed robustness and accuracy of the predictions is even cited as an argument to tout the merits of DNA motif finding software. Software developers' reluctance to endorse any tool not meeting the (purely statistical and thus theoretical) reliability criteria results in programs that have built-in high threshold scores or *P*-values, limiting the output to the very best matches. However, the score obtained by a sequence does not accurately reflect the affinity of a TF for its predicted *cis* sites, but in fact only highlights the similarity of this sequence with the training set. Preventing users from defining the threshold is a way to prevent the 'misuse' of the prediction software, and for statisticians misuse means producing an output list full of false positive hits. The problem is that this arbitrary constraint is only based on probabilistic standards depending only on the training set of sequences that have been used to generate the PWM instead of being supported by biochemical values (see below the paragraph on the 'historical background').

Software users aim for exhaustive prediction outputs

In contrast, simply predicting the straightforward, basic and expectable binding sites of a TF is not generally the ambition of software users, often biologists, who instead aim for an exhaustive regulon prediction output, i.e. an output that additionally provides non-obvious positive hits. For the biologist, a predicted DNA-binding site of a TF (highly reliable or not) simply represents a statistical information requiring further experimental validation through classical (one gene at a time) or high-throughput experimental techniques. Weakly-bound sites of a TF most often fall into this 'non-obvious and unreliable' category although they are just as important as strong binding sites in the route of unveiling the molecular mechanisms controlling the triggering of a biological process. For instance, in many inducible systems key genes are expressed at a basal level that allows the organism to possess a sufficient amount of proteins to sense the activating signal once it is present in the environment. This basal expression level necessary for sensing the regulon elicitor is only possible if the targeted DNA sequence of a transcriptional repressor has undergone a series of mutations, weakening the DNA-TF interaction and thus possibly escaping the threshold score fixed by the prediction software. In addition, when the expression of a gene is controlled by many different TFs, each TF must include its own binding site within a gene's upstream region, which is limited in size. The multiplicity of TFs controlling a single gene often signifies that the expression of the target gene is elicited by many different environmental signals. When several TFs bind neighbouring or overlapping sequences this necessarily implies the evolution of non-discriminatory mutations, allowing the different binding sites to coexist.

Besides the restrictions imposed by software creators, biological and technical causes are also responsible for favouring the discovery of the strongest binding sites of a TF. Indeed, independently of the algorithm used by the DNA motif search software, the prediction output almost entirely depends on the training set of sequences used to generate the PWM. This set of sequences is in turn dependent on the 'historical background' of the selected TF (i.e. the earlier investigations that identified the *cis*-acting elements already discovered). In general, DNA sequences most strongly bound by the regulatory protein (those that best match with the preferred sequence of the TF), are the sequences that tend to be identified first. This is in part due to the fact that the strongest interactions are technically easier to detect than weaker ones, and that the principal target genes of a TF (those that requires strict expression control) are often located in the neighbouring region of the TF and are also conserved between species, which further facilitates their discovery. This natural inclination for the *cis*-acting elements most strongly bound by a TF to be discovered first has a strong impact on the efficiency of regulon predictions, as the earliest PWMs are biased towards the discovery of highly reliable sequences.

Overall, the software user has numerous reasons to not accept the threshold or cut-off scores fixed by software developers, legitimately concerned with providing outputs with a limited number of false positive hits.

Examples of weakly-bound sites not detected by web tools that fix restrictive threshold scores

To illustrate that favouring reliable instead of exhaustive outputs has little biological meaning we decided to demonstrate that a series of weakly bound *cis*-acting elements of a well-studied TF could never have been discovered using the threshold scores imposed by some of the most popular regulon prediction web tools. The examples chosen originate from our investigations into the regulon of the N-acetylglucosamine (GlcNAc) utilisation regulator DasR in *Streptomyces coelicolor*[3]. Chronologically, all first target genes experimentally demonstrated to be controlled by DasR were genes that revolve around the catabolism of GlcNAc and its polymer chitin[3-6] (Figure 1). As often happens for global or pleiotropic regulators, the first sequences that we discovered to be bound by DasR (referred to as "*dre*" for DasR responsive element), were the best ones, i.e. those best matching the *dre* palindromic sequence ACTGGTCTAGACCAGT. Indeed, as presented in Table 1, the consensus sequences deduced from *dres* upstream of genes involved in chitin and GlcNAc utilisation both exhibited very high scores, with respectively one and two mismatches compared with the perfect 16-bp palindromic *dre*.

Table 1.
Sequences and scores of DasR responsive elements

Gene(s)	*dre*	Score	SD
Palindromic *dre*	ACTGGTCTAGACCAGT	16.90	na
GlcNAc genes	AgTGGTCTAGACCAcT	13.70*	1.31
chi genes	ACTGGTCTAGACCAaT	12.02*	1.86
*dmdR*1	tgcGGTCTgGACCAGT	9.73	na
redZ	AgTGGTtTccACCtca	5.95	na
*act*II-4	tgTtGacTAGgCCtGT	2.85	na

* The presented score is the mean of the scores of the best *dres* identified upstream of *chi*, and GlcNAc associated genes. na = not applicable. SD = standard deviation. Lower case letters indicate nucleotides that differs from palindromic *dre*.

In order to identify genes controlled by DasR beyond the predictable GlcNAc and chitin utilisation systems, we created in 2007 our own DNA motif screening tool named PREDetector (Prokaryotic Regulatory Elements Detector[7]), motivated (and frustrated) by the unavailability of threshold-free regulon prediction software products. The threshold score that is most often recommended by regulon prediction web tools is the weakest score of the sequences used to generate the PWM (hereafter referred to as "T_{MS}", for threshold based on the PWM minimal score). As, in our case, the first 15 experimentally validated *dres* were used to generate the PWM, this primary matrix was naturally biased for the identification of highly reliable *dres*. Indeed, at T_{MS}, >90% of *chi* and GlcNAc-related genes happened to possess at least one *dre*, confirming that the highly predictable target genes of DasR (based on their biological function) are also the genes with *cis*-acting sequences that best match the palindromic *dre* sequence (Figure 1 and Table 1).

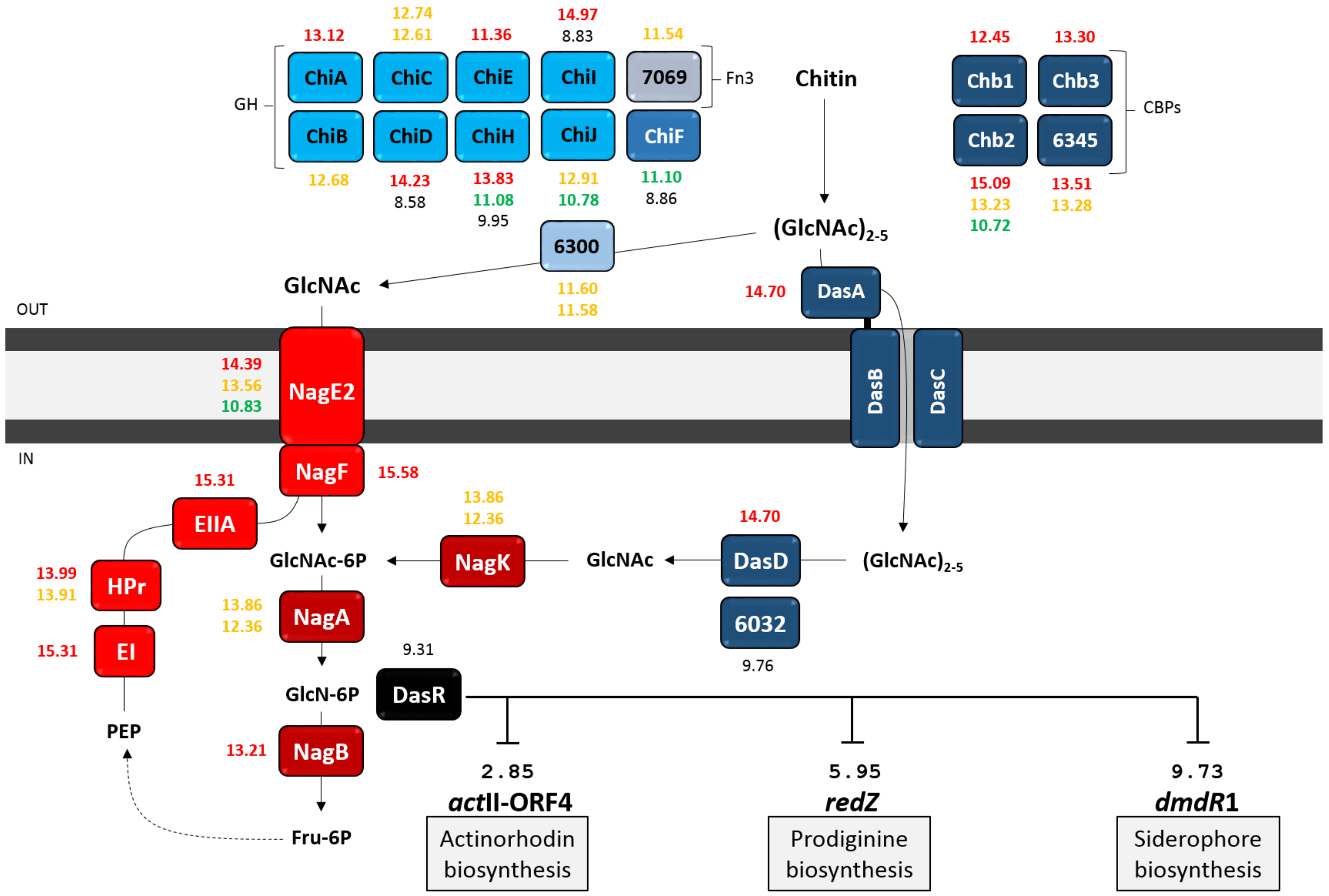

Figure 1. Scores of *dres* associated with proteins of the DasR regulon in *Streptomyces coelicolor*. N-acetylglucosamine (GlcNAc) originates from chitin, chitoolichosacharides $[(GlcNAc)_{3\text{-}5}]$, and *N*-*N*'-diacetylchitobiose $[(GlcNAc)_2]$ hydrolysis by the chitinolytic system (proteins in blue). Elements for uptake, phosphorylation and catabolism of GlcNAc are highlighted in red. Numbers next to proteins indicate the score obtained by their respective *dres* using the PWM 'DasR2008' generated via the PREDetector software[7]. *dres* used to generate the PWM are highlighted in red. *dres* found at T_{MS} (cut-off score = 11.36) are highlighted in orange. *dres* found above and below $T_{50\%}$ (cut-off score = 10.32) are highlighted in green, and black colors, respectively. For full description of the function of proteins see Liao and Rigali et al.[8].

When running PREDetector without threshold score restrictions, the program suggested possible *dres*, though at very low scores, upstream of TFs associated with biological processes previously unrelated to chitin/GlcNAc catabolism (Figure 2A and Table 1). Our particular interest in finding environmental signals that could possibly control the triggering of secondary metabolite production raised our attention on three predicted *dres*: (i) a *dre* identified upstream of *redZ* (dre^{redZ}), a TF that activates expression of the pathway-specific activator *redD* of prodiginines production; (ii) the *dre* predicted upstream of *act*II-ORF4 ($dre^{act\text{II-4}}$), the pathway-specific activator of actinorhodin production; and (iii) the putative *dre* upstream of *dmdR*1 (dre^{dmdR1}), encoding for the repressor of siderophores, desferrioxamines and coelichelin biosynthesis. As Figure 2A illustrates, $dre^{act\text{II-4}}$, dre^{redZ}, and dre^{dmdR1} fell well below T_{MS}. Another type of cut-off score was proposed by Tan et al.[9] who suggested fixing the threshold score for weak sites at which greater than half of all sites are located upstream of transcription units (hereafter referred to as "$T_{50\%}$"). This arbitrary threshold is based on the naturally strong biased location of *cis*-acting elements in genes' upstream regions (>90%) rather than within genes' coding sequences. A cut-off of $T_{50\%}$ can thus be expected to generate an output list containing a large proportion of false positive hits. Here again, scores obtained for dre^{dmdR1}, $dre^{act\text{II-4}}$, dre^{redZ}, are still below $T_{50\%}$ (Figure 2A). Even worse, $dre^{act\text{II-4}}$ and dre^{redZ}, have scores at which almost all sites are located within transcription units and below the threshold of random localisation (hereafter referred to as "T_{RL}"). We defined the T_{RL} as suggested by Tan et al[9] based on the proportion non-coding and coding regions in the microorganisms where the DNA motif screening is performed. The *S. coelicolor* genome has ~11% of non-coding regions. Thus, sequences with random localisation, i.e. not biased for either coding or non-coding sequences, will occur at scores where hits are located ~11% of the time in non-coding regions and ~89% within transcription units. $dre^{act\text{II-4}}$ and dre^{redZ} have scores at which 91 and 89% of all sites are located in transcription units (Figure 2A), clearly indicating that these sites belong to the background noise of the DNA motif screening process. Notwithstanding these ‘worst case scenario’ statistics (scores below T_{RL}), we decided to further investigate these hits and the binding of DasR to $dre^{actII\text{-}4}$, dre^{redZ}, and dre^{dmdR1} was demonstrated by electromobility gel shift assays[10, 11], while the repressing activity of DasR on the expression of the latter genes was demonstrated by RT-PCR[10, 11]. *In vivo* binding of DasR to $dre^{actII\text{-}4}$, and dre^{redZ}, was also recently observed through Chromatin ImmunoPrecipitation-sequencing. The use of a threshold-free regulon prediction software thus allowed, for the first time, the identification of complete signalling pathways from elicitor sensing to secondary metabolite production in the industrially important ‘antibiotic makers’ *Streptomyces* species[12].

To prove that the inability to identify weak sites was due to the restrictive screening parameters fixed by software developers, we repeated our prediction of the DasR regulon by means of other algorithms used by some of the most popular regulon prediction web tools. As obtained with the algorithm used in PREDetector (Figure 2A), none of the other algorithms allowed to detect $dre^{actII\text{-}4}$, dre^{redZ}, and dre^{dmdR1} at T_{MS} or at $T_{50\%}$. $dre^{actII\text{-}4}$ and dre^{redZ} are always detected at scores around or below T_{RL} (Figure 2). The threshold set by default by the DNA motif finding web tools (T* or $T^{\ddagger}$ in Figure 2) is often above $T_{50\%}$. The FIMO[13] (Find Individual Motif Occurrences) software from the MEME suite toolkit is the only web tool tested that fixed its default output threshold low enough (even below T_{RL}) to detect dre^{redZ}, and dre^{dmdR1} (Figure 2C). Importantly, FIMO, likewise PREDetector, also allows user to considerably modify the output threshold set by default (marked as $T^{\ddagger}$ in Figure 2A and C). Particularly interesting is the comparison of the drastically different results obtained by the two algorithms proposed by the FITBAR software[14]. When predictions are performed via the entropy-weighted position-specific scoring matrices the threshold set by default is below $T_{50\%}$ and includes dre^{dmdR1} in the output list (Figure 2E), while using the log-odds algorithm results in predictions which do not even include the sequences from the training set (Figure 2F), as also observed with the threshold set by default by the PRODORIC web-tool (Figure 2B).

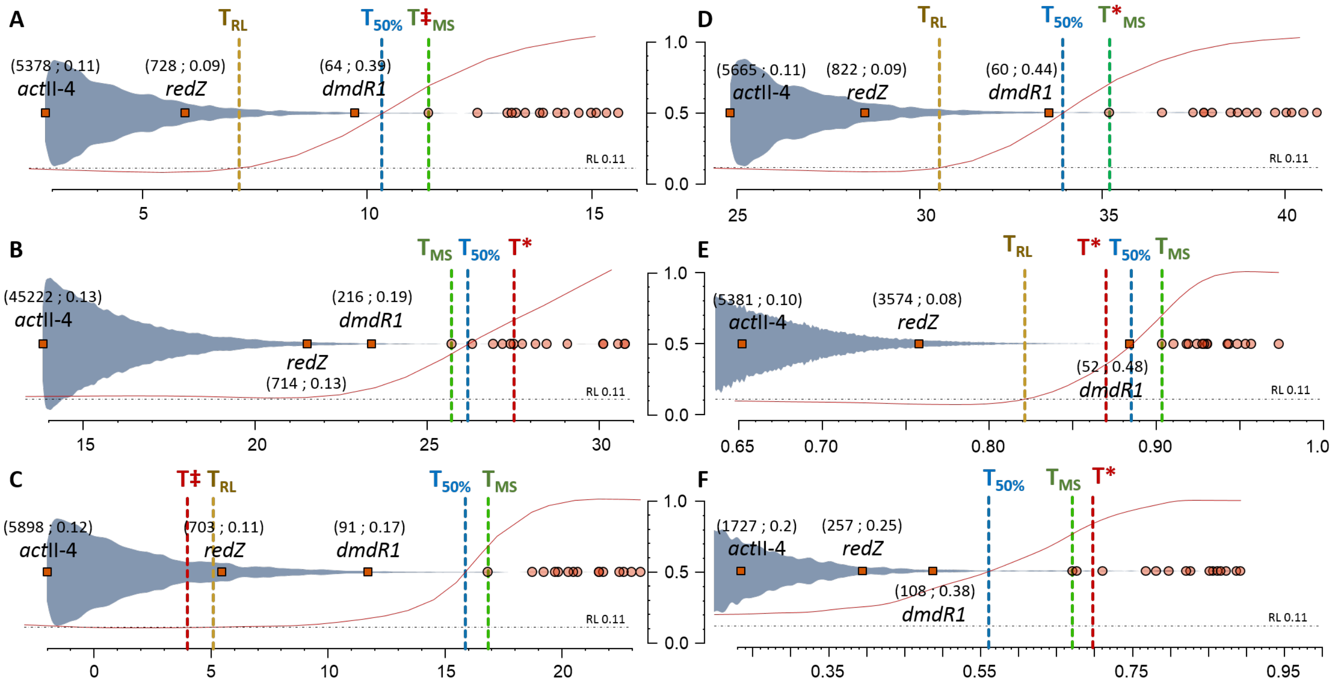

Figure 2. Position of dre^{dmdR1}, $dre^{actII\text{-}4}$, and dre^{redZ} according to different arbitrary thresholds and based on scores obtained via different algorithms. A. PREDetector (*pi* based on training set). **B.** PRODORIC[15]. **C.** FIMO[13]. **D.** RegPrecise[16]. **E.** FITBAR[14] (entropy-weighted position-specific scoring matrices). **F.** FITBAR (log-odds). X axis = arbitrary scores given according to the different algorithms. Y axis = percentage of hits in a certain region (intergenic versus coding sequences). Red line = % of hits in a certain region according to the score. Blue beanplot = total number of hits according to the score. Circles = positions of sequences used to generate the PWM. T* or $T^{\ddagger}$ = Rigid or flexible thresholds fixed by default by the software, respectively. Numbers between brackets = numbers of total hits found before dre^{dmdR1}, $dre^{actII\text{-}4}$, and dre^{redZ} and percentage of hits in a certain region at their respective score. Algorithms that do not allow free modification of the threshold set by default were re-coded in R[17] in order to create a consistent framework for comparison.

Conclusions

In their sarcastic commentary ‘How Not to Be a Bioinformatician’, Manuel Corpas et al. wrote that ‘to blindly believe in the predictions given, P-values or statistics’ is one of the top ten disastrous practices in the bioinformatics field[18]. Sadly, this is often a common practice among software developers, who give priority to the statistical ‘robustness’ of their program by suggesting, or even imposing, stringent cut-offs whose meaning mistakenly becomes “*absolute truth above, absolute falsehood below*[18]”. As long as TF-binding sites (TFBS) prediction algorithms will mostly focus on sequence similarity - which only imperfectly accounts for the multiple and complex rules governing TF-TFBS interactions - scoring and associated statistics will fail to optimally help software users. We illustrated through three examples why the choice between reliable or exhaustive DNA motif prediction outputs is a dilemma that has no real biological meaning. The discovery of the DasR and GlcNAc-mediated control of secondary metabolite production had now become a paradigm for actinomycetes developmental studies[10, 12]. These examples should convince software users that many biologically relevant hits are hidden below the arbitrary fixed threshold scores and that spending time analysing a long output list is often rewarded. We thus encourage software users to privilege the utilization of web tools that leave them the opportunity to lower the prediction cut-off score set by default, which will obviously result in a longer list of putative *cis*-acting sites, full of false positive hits. For large eukaryotic genomes, the length of this list and the total number of potential hits may be quite large and hence more difficult to go over them in detail to prioritise them. However, biologists should not be scared of 'making mistakes', which in this case would be to decide to start investigations on a false positive hit. Finding unexpected regulatory connections implies taking risks and is time consuming. In any case, the study of an *in silico* predicted *cis* site that is eventually experimentally demonstrated not to occur *in vivo* is not necessarily considered a loss of time as such an investigation constitutes a negative control that strengthens the credibility of the positive interactions previously identified.

The challenge for bioinformaticians is to provide tools that would minimize risks and therefore the main improvements of next generation regulon prediction web-tools should aim at helping the user in finding which hits are potentially worth to investigate in the infinite list of putative candidates. Nevertheless, whatever how improved will be the updated versions of regulon predicting tools, the decision to start investigations on a possible TF-binding site that escapes the statistical criteria of reliability must only be the biologist's own decision, based on his/her knowledge, enthusiasm, and curiosity.

Acknowledgements

SR is a FRS-FNRS research associate. This work is supported in part by the Belgian program of Interuniversity Attraction Poles initiated by the Federal Office for Scientific Technical and Cultural Affairs (PAI no. P7/44). We are thankful to Charles Michael Kittridge Thompson IV (Threshold apprehension, Cooking vinyl) and to Ernie Freeman (Limbo Dance Party, Liberty) for inspiring us the graphical abstract of our manuscript.

Notes

[a]Centre for Protein Engineering, University of Liège, Institut de Chimie B6a, B-4000 Liège, Belgium.
[b]Laboratory of Plant Physiology, PhytoSYSTEMS, University of Liège, B-4000 Liège, Belgium
[c]Corresponding author, srigali@ulg.ac.be

References

1. J. Mrazek, *Brief Bioinform*, 2009, 10, 525-536.
2. G. D. Stormo, *Bioinformatics*, 2000, 16, 16-23.
3. S. Rigali, H. Nothaft, E. E. Noens, M. Schlicht, S. Colson, M. Muller, B. Joris, H. K. Koerten, D. A. Hopwood, F. Titgemeyer and G. P. van Wezel, *Mol Microbiol*, 2006, 61, 1237-1251.
4. S. Colson, G. P. van Wezel, M. Craig, E. E. Noens, H. Nothaft, A. M. Mommaas, F. Titgemeyer, B. Joris and S. Rigali, *Microbiology*, 2008, 154, 373-382.
5. S. Rigali, M. Schlicht, P. Hoskisson, H. Nothaft, M. Merzbacher, B. Joris and F. Titgemeyer, *Nucleic Acids Res*, 2004, 32, 3418-3426.
6. S. Colson, J. Stephan, T. Hertrich, A. Saito, G. P. van Wezel, F. Titgemeyer and S. Rigali, *J Mol Microbiol Biotechnol*, 2007, 12, 60-66.
7. S. Hiard, R. Maree, S. Colson, P. A. Hoskisson, F. Titgemeyer, G. P. van Wezel, B. Joris, L. Wehenkel and S. Rigali, *Biochem Biophys Res Commun*, 2007, 357, 861-864.
8. C. Liao, S. Rigali, C. L. Cassani, E. M. Saldana, L. K. Nielsen and B. C. Ye, *Microbiology*, 2014.
9. K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides and G. D. Stormo, *Genome Res*, 2001, 11, 566-584.
10. M. Craig, S. Lambert, S. Jourdan, E. Tenconi, S. Colson, M. Maciejewska, M. Ongena, J. F. Martin, G. van Wezel and S. Rigali, *Environ Microbiol Rep*, 2012, 4, 512-521.
11. S. Rigali, F. Titgemeyer, S. Barends, S. Mulder, A. W. Thomae, D. A. Hopwood and G. P. van Wezel, *EMBO Rep*, 2008, 9, 670-675.
12. S. Molloy, *Nature Reviews Microbiology*, 2008, 6.
13. C. E. Grant, T. L. Bailey and W. S. Noble, *Bioinformatics*, 2011, 27, 1017-1018.
14. J. Oberto, *BMC Bioinformatics*, 2010, 11, 554.
15. R. Munch, K. Hiller, A. Grote, M. Scheer, J. Klein, M. Schobert and D. Jahn, *Bioinformatics*, 2005, 21, 4187-4189.
16. P. S. Novichkov, A. E. Kazakov, D. A. Ravcheev, S. A. Leyn, G. Y. Kovaleva, R. A. Sutormin, M. D. Kazanov, W. Riehl, A. P. Arkin, I. Dubchak and D. A. Rodionov, *BMC Genomics*, 2013, 14, 745.
17. R. B. Dessau and C. B. Pipper, *Ugeskr Laeger*, 2008, 170, 328-330.
18. M. Corpas, S. Fatumo and R. Schneider, *Source Code Biol Med*, 2012, 7, 3.