Genome-specific higher-order background models to improve motif detection

Kathleen Marchal[1], Gert Thijs[1], Sigrid De Keersmaecker[2], Pieter Monsieurs[1], Bart De Moor[1] and Jos Vanderleyden[2]

[1]ESAT SISTA-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium
[2]Centre of Microbial and Plant Genetics, K.U.Leuven, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium

Motif detection based on Gibbs sampling is a common procedure used to retrieve regulatory motifs *in silico*. Using a species-specific background model was previously shown to increase the robustness of the algorithm. Here, we demonstrate that selecting a non-species-adapted background model can have an adverse effect on the results of motif detection. The large differences in the average nucleotide composition of prokaryotic sequences exacerbate the problem of exchanging background models. Therefore, we have developed complex background models for all prokaryotic species with available genome sequences.

DNA motifs are short patterns of DNA. In the promoter regions of genes, motifs constitute the recognition site of transcriptional regulators, and thus reflect the underlying transcriptional networks active at the cellular level. Elucidating such regulatory elements will help to unravel these networks and gain insights into global cellular regulation.

Motif-detection strategies involve searching for DNA patterns that are present more frequently in a set of related sequences than in a set of unrelated sequences. 'Related sequences' here refers to genes that are co-expressed or co-regulated and are therefore expected to share similar conserved regulatory motifs. Such co-expressed genes can be identified using high-throughput gene-expression profiling experiments [1,2]. Alternatively, instead of co-regulated genes, intergenic sequences of orthologous genes can also constitute a valuable dataset for motif detection [3,4]. In this case, motif detection is referred to as phylogenetic footprinting. If selection pressure tends to conserve DNA patterns in the intergenic regions of homologous genes in related species, such DNA patterns can be expected to be biologically relevant and to reflect a conserved ancestral mode of regulation.

Motif-detection algorithms such as Gibbs sampling identify conserved patterns based solely on statistical properties, that is, no prior information on what the motif should look like is required [5]. Currently, several motif-detection algorithms based on Gibbs sampling are freely accessible (e.g. Bioprospector [6], AlignACE [7], Motif Sampler [8] and ANN-spec [9]). Each of these algorithms, although based on the same Gibbs sampling strategy, differs slightly in the way it is implemented. Several studies [3,4,10,11] have already demonstrated the usefulness of these methods for bacterial motif detection and phylogenetic footprinting.

However, a major drawback of these statistical *in silico* motif-detection approaches is their sensitivity to the presence of 'noise'. Noise, in the context of motif detection, corresponds to areas of sequence in the dataset that do not contain the consensus pattern. They originate either from genes not containing the over-represented motif in their promoter region or from genes for which the length of the intergenic sequence is large relative to the length of the motif. A Gibbs sampler always makes a trade-off between the degree of conservation of a retrieved pattern and the frequency of occurrence of this pattern (i.e. the higher the number of hits, the more statistically relevant the motif). Therefore, if a well-conserved motif is present in only a limited number of sequences, the algorithm will preferentially select a less conserved but more frequent motif, which often corresponds to a pattern that is over-represented because of the organism's general nucleotide composition (i.e. background). The operon-like organization of genes in bacterial species increases the problem: in a set of co-expressed genes only small subsets (the first genes of the operon) are expected to contain the motif and the intergenic regions of the other genes contribute to the noise.

One way to improve the robustness of the algorithm to noise (i.e. lower the variability of the outcome of the algorithm) is to use an independent, species-specific higher-order background model. A background model is a mathematical representation of the areas of the sequence that do not contain motifs. The better the representation of the background, the higher the efficiency of detecting true positive motifs in the presence of noise [12]. Improved background models, mainly for *Saccharomyces cerevisiae*, are also implemented in BioProspector [6] and ANNspec [9]. The use of a species-specific background model means the algorithm distinguishes better between patterns specific for the set of co-expressed genes under study versus patterns that also occur frequently in sets of unrelated sequences from the same genome.

Species-specific higher-order background models

To facilitate the use of Gibbs sampling in prokaryotes, higher-order background models for all species with

Corresponding author: Kathleen Marchal (Kathleen.Marchal@esat.kuleuven.ac.be).

 0966-842X/03/$ - see front matter doi:10.1016/S0966-842X(02)00030-6

genome sequences available in GenBank (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) were constructed. These higher-order background models (Markov models) can be used in combination with the Motif Sampler [8] and are available at http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html. To construct the background models, intergenic regions for each completely sequenced genome were selected using the modules of INCLUSive [13]. Because their respective nucleotide compositions can differ significantly, separate models were built for plasmids and genome sequences. Details on how background models were calculated are displayed as supplementary information on http://www.esat.kuleuven.ac.be/~thijs/help/help_background.html.

The oligonucleotide composition of the intergenic regions, summarized by a species-specific vector containing the transition probabilities [12], was used to compute the distance between the different species. The relationship between the species inferred from this measure of distance is summarized in a hierarchical tree (see supplementary information). This tree partially reflects the generally accepted phylogenetic relationships and can be used as a guideline to select background models that can be interchanged between microorganisms. Figure 1 gives a visual representation of the relationship between the 3rd-order transition probabilities [12] of two species for which, according to the hierarchical tree, the background models are similar – *Escherichia coli* K12 and *Salmonella typhimurium* – and two species with very different background compositions – *E. coli* K12 and *Streptomyces coelicolor* A32.

The well-known σ^{54} consensus motif [14] was searched for in a dataset of *E. coli* genes and a corresponding set of *Pseudomonas aeruginosa* genes (Table 1) to illustrate the influence using non-species-specific background models has on the efficacy of the algorithm. The bacterial alternative sigma factor σ^{54} (or RpoN) recognizes a specific $-12/-24$-type promoter [15]. It controls several ancillary processes such as assimilation of ammonia, hydrogen uptake, nitrogen fixation, flagellar assembly and arginine catabolism (see [14] and [16] for reviews). Because σ^{54} is a widely distributed regulatory factor, its recognition motif is conserved in distantly related bacterial species with largely distinct background compositions (such as *E. coli* and *P. aeruginosa*). Moreover, as can be derived from its consensus sequence (5′-TGGCACG-N4-TTGCWN-3′) [15], the motif contains some non-informative positions and it is thus not a trivial task to retrieve this motif using motif detection. The σ^{54} consensus sequence thus optimally suits our purpose of illustrating the influence different background models can have on motif retrieval.

For both the *E. coli* and *P. aeruginosa* dataset, we tested the influence of four background models with very different compositions (Table 2) Because the result of a motif-sampling test also depends on other parameters, such as motif length and order of the background model, the tests were repeated for motif lengths of 7 bp and 17 bp and different higher-order background models. Results are displayed for 0th- (i.e. single-nucleotide frequency) and 3rd-order background models only. Motif detection based on Gibbs sampling is a stochastic procedure, which means that running the algorithm with exactly the same parameter settings and input data does not necessarily retrieve the same motifs. The number of potential motifs that can be detected is huge and most of the local optima correspond to coincidental local alignments that are not true motifs. The power of Gibbs sampling is that it can escape from such optima and search for a motif with a higher score. Retrieving a motif by Gibbs sampling implies running the algorithm repeatedly with the same parameter settings and calculating the statistics of the outcome. Indeed, the better a solution in a given dataset, and thus the higher the number of instances and the stronger its conservation, the more frequent it will be retrieved over different runs. The number of times a motif is retrieved on 100 runs of the algorithm is therefore a

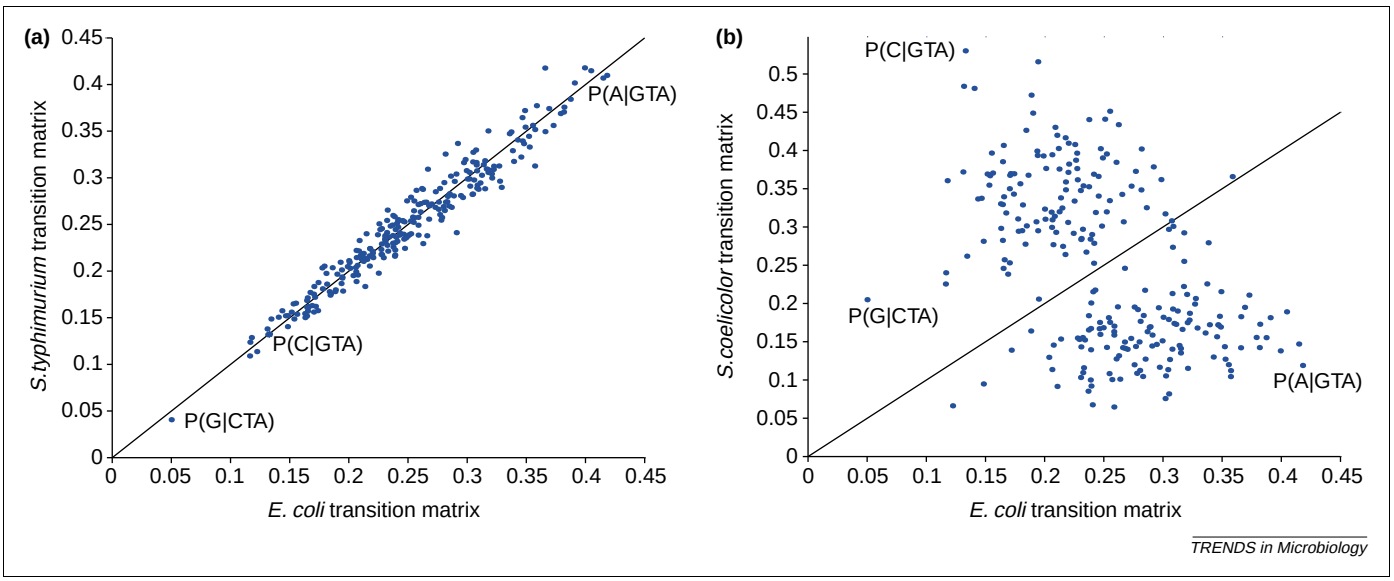

Fig. 1. Comparison of 3rd-order transition probabilities of two different genomes. Each dot corresponds to an entry in the transition matrix, which gives the probability of a nucleotide given the preceding trimer. (a) Plot of the transition probabilities of genomes with similar background composition (*Escherichia coli* K12 and *Salmonella typhimurium*). The similarity is expressed by the straight line, reflecting the one-to-one relationship. (b) Plot of the transition probabilities of genomes with different background composition (*E. coli* K12 and *Streptomyces coelicolor* A32). The plot clearly reflects the distinct background composition in both genomes.

Table 1. Overview of the *Escherichia coli* and *Pseudomonas aeruginosa* datasets

Gene	Description[a]	Length (bp)[b]	Source[c]
***Escherichia coli* K12**			
argT	Arginine-, ornithine-binding periplasmic protein	267	1
ygjC	Probable ornithine aminotransferase	308	1
hisJ	Histidine-binding periplasmic protein of high-affinity histidine transport system	222	1
atoD	Acetyl-CoA:acetoacetyl-CoA transferase alpha subunit	197	1
fdhF	Selenocysteine selenopolypeptide subunit of formate dehydrogenase H	199	1
glnA	Glutamine synthetase	374	1
glnH	Periplasmic glutamine-binding protein; permease	405	1
glnK	Nitrogen regulatory protein P-II 2	182	1
hycA	Alternate gene name *hevA*; transcriptional repression of *hyc* and *hyp* operons	213	1
hypA	Pleiotrophic effects on 3 hydrogenase isozymes	213	1
hydN	Involved in electron transport from formate to hydrogen, Fe-S centres	150	1
hydH	Sensor kinase for HydG, hydrogenase 3 activity	98	1
prpB	Putative phosphonomutase 2	240	1
pspA	Phage shock protein, inner membrane protein	153	1
rtcB	Formerly designated *yhgL* orf, hypothetical protein	190	1
Pseudomonas aeruginosa			
aotJ	PA0888 arginine*ornithine binding protein AotJ	542	4
arcD	PA5170 arginine*ornithine antiporter	797	4
PA5530	Probable MFS dicarboxylate transporter	317	4
glnA	PA5119 glutamine synthetase	338	3
glnK	PA5288 nitrogen regulatory protein P-II 2	441	3
fleS	PA1098 two-component sensor	114	2
pilT	PA0395 twitching motility protein PilT	216	4
pilA	PA4525 type 4 fimbrial precursor PilA	233	4
fliE	PA1100 flagellar hook-basal body complex protein FliE	248	4
flgB	PA1077 flagellar basal-body rod protein FlgB	254	4
algC	PA5322 phosphomannomutase AlgC	1253	2
algD	PA3540 GDP-mannose 6-dehydrogenase AlgD	903	2
oprE	PA0291 outer membrane porin OprE precursor	614	2
cpg2	PA2787 carboxypeptidase G2 precursor	68	2
PA2128	PA2128 probable fimbrial protein	854	4
RhlA	PA3479 rhamnosyltransferase chain A	425	2

[a]The functional annotation of each gene, derived from GenBank [17].
[b]The length of the intergenic region used for motif detection.
[c]The source of the information that was used to select the genes: (1), the *E. coli* set was compiled based on [16] and contains 15 experimentally confirmed σ^{54}-dependent promoters; (2), genes that contained a σ^{54} site in the -12/-24 region upstream of the transcription start site as predicted by previous studies [14]; (3), genes for which a genomic screen of *P. aeruginosa* with the σ^{54} motif model of *E. coli* showed the presence of a putative σ^{54} consensus and that were orthologues of the verified *E. coli* targets; and (4) genes for which a genomic screen of *P. aeruginosa* with the σ^{54} motif model of *E. coli* showed the presence of a putative σ^{54} consensus and had a function related to known σ^{54} targets (i.e. genes involved in flagellar assembly, arginine catabolism and dicarboxylic acid transport [14]).

measurement of the stability of the motif and expresses a confidence in its prediction. The motifs obtained after 100 runs of the algorithm with a given parameter setting are ranked according to their log-likelihood score [8]. The log-likelihood is the score that, in our opinion, best summarizes the specificities of a true motif. A log-likelihood score will depend on the degree of conservation of the motif, a characteristic also reflected by a high consensus score, and on the number of instances of that motif in the dataset.

Table 2 shows that for both motif lengths, 7 bp and 17 bp, using a species-specific background leads to the retrieval of the σ^{54} consensus as one of the top scoring motifs. Its relatively high log-likelihood score can be attributed to a relatively high consensus score and a reasonable number of occurrences in the dataset (because we assume that the motif occurs once in each sequence, 15 and 18 hits are expected for the *E. coli* and *P. aeruginosa* datasets, respectively). Using a background model of lower order decreases the performance of the algorithm, which is most clear for the results on the *P. aeruginosa* dataset. The σ^{54} motif is still retrieved when using an appropriate species-specific background model but is no longer the motif with the highest score (Table 2).

Using a non-species-specific background model (i.e. the 'wrong' background model) will generally prohibit retrieval of the true motif and result in the detection of highly degenerated motifs. High-ranking motifs retrieved using the wrong background model have a high log-likelihood score but this is the result of an extremely low consensus score [8] and an unreasonably high number of instances (Table 2). The use of a GC-rich background model [(e.g. *S. coelicolor* (71% GC) and *P. aeruginosa* (61% GC)] in an AT-rich organism usually promotes the retrieval of AT-rich degenerated motifs (Table 2) while the opposite is true for the use of AT-rich background models (e.g. *Listeria monocytogenes*, 66% AT) in GC-rich organisms (Table 2). When using a completely wrong background model, lowering the order is a logical option because a lower-order background model captures less of the species-specific sequence complexity. In our test example, using a non-species-specific background of 0th order instead of a 3rd-order model did not improve detection of the true σ^{54} motif (Table 2). However, note also that retrieving the motifs becomes more difficult when using a lower-order background model (owing to the presence of more false positives). Therefore, for a background model of an organism for which the nucleotide composition is expected

Table 2. The influence of four different background models on detection of the σ^{54} motif

Background model[a]	Consensus	Log-likelihood score		Consensus score		Rank[b]	No. instances[c]	No. hits/100[d]
	Highest scoring	Highest scoring	σ^{54}	Highest scoring	σ^{54}	σ^{54}	Highest scoring	σ^{54}
***E. coli* dataset**								
Order 3, length 17								
E. coli	TGGCACrAywmnTGCAT	167.48	167.48	1.0969	1.0969	1	13	37
L. monocytogenes	TGGCACrAywmnTGCAT	179.44	179.44	1.0969	1.0969	1	13	36
P. aeruginosa	wnwwwnAnnnmATwATw	151.02	–	0.5118	–	–	63	0
S. coelicolor	wwwwnynkyrmwnnwnw	200.21	–	0.2846	–	–	132	0
Order 0, length 17								
E. coli	TGGCACrAywmnTGCAT	177.17	177.17	1.0969	1.0969	1	13	23
L. monocytogenes	TGGCACrAnwnnTGCwT	195.17	195.17	1.076	1.076	1	14	57
P. aeruginosa	wwwwwnAksrmAnnwww	159.87	–	0.4539	–	–	79	0
S. coelicolor	wwwwwynnnnmnwnwww	193.54	–	0.2862	–	–	132	0
Order 3, length 7								
E. coli	TGGCACr	121.37	121.37	1.5467	1.5467	1	17	23
L. monocytogenes	TGGCACr	132.2	132.2	1.5015	1.5015	1	19	53
P. aeruginosa	TwmTTAA	122.71	–	1.1472	–	–	47	0
S. coelicolor	nwnwnAw	145.64	–	0.7037	–	–	138	0
Order 0, length 7								
E. coli	TGGCACr	124.16	124.16	1.5467	1.5467	1	17	21
L. monocytogenes	TGGCACr	131.43	131.43	1.5467	1.5467	1	17	52
P. aeruginosa	wTAAmAr	122.36	–	1.0184	–	–	63	0
S. coelicolor	ATwwTnA	139.4	–	0.8052	–	–	124	0
***P. aeruginosa* dataset**								
Order 3, length 17								
E. coli	yCGsGsCsknCssnnG	152.86	–	0.5681	–	–	83	0
L. monocytogenes	CGsnGmnsnnnnssCss	194.07	–	0.3646	–	–	192	0
P. aeruginosa	nTGGCACGsnwnTTGCT	142.92	142.92	1.0869	1.0869	1	11	37
S. coelicolor	nwwnkrnwyryynAwnn	178.09	–	0.3791	–	–	71	0
Order 0, length 17								
E. coli	ssnnGsCnnnsnCsssn	167.62	–	0.4947	–	–	121	0
L. monocytogenes	ssnnsnsnsnnnnsnss	204.6	–	0.3313	–	–	211	0
P. aeruginosa	TTsywnnynyTTTGnnn	134.1	124.65	0.7565	1.2322	4	17 (8)	14
S. coelicolor	nwwnTnnwnrnTwnTnr	158.52	–	0.42321	–	–	60	0
Order 3, length 7								
E. coli	CGCGsCk	126.41	–	1.3551	–	–	47	0
L. monocytogenes	sCsnGsC	153.9	–	1.1023	–	–	125	0
P. aeruginosa	wTTGGCA	111.43	111.43	1.4142	1.4142	1	16	15
S. coelicolor	nTkmwAw	121.08	–	0.0808	–	–	66	0
Order 0, length 7								
E. coli	ssCGGCs	140.69	–	1.2654	–	–	82	0
L. monocytogenes	sCsCGsC	162.26	–	1.0372	–	–	154	0
P. aeruginosa	TTTTnCk	110.66	109.06	1.2836	1.416	2	23 (18)	4
S. coelicolor	WmTwwTT	118.46	–	0.8946	–	–	56	0

[a]The parameter settings were as follows: motif length, 17 bp and 7 bp; order of background model, 3 and 0; maximal number of occurrences for a motif, 1; number of distinct motifs, 1. For each parameter setting, 100 runs of the algorithm were performed.
[b]The motif rank is the position of the motif among the highest scoring motifs according to their log likelihood; – indicates no σ^{54} motif was detected.
[c]Expresses the number of occurences of the highest scoring motif in the dataset (if the highest scoring motif differs from the σ^{54} motif, the number of instances of the σ^{54} motif is indicated in brackets).
[d]Expresses the number of times the σ^{54} motif was recorded on 100 runs of the algorithm.

to be similar to that of the species of interest, using this background model with a higher order might still be more appropriate.

Using the wrong background model can occasionally retrieve the motif of interest such as we observed in our example: the *L. monocytogenes* background model can perform as well or might even slightly outperform the species-specific *E. coli* model in retrieving the *E. coli* σ^{54} motif. Whether or not this will occur depends to a great extent on the specificities of the motif searched for and its relation to the background model, factors that can only be estimated retrospectively. Therefore, it is advisable whenever possible to use a species-specific higher-order background model or a higher-order background model of a related species.

Conclusions

As the number of genome-wide high-throughput expression profiling experiments steadily increases and microbiologists rely more and more on systems biology to unravel regulatory pathways, the importance of motif detection as an *in silico* method will increase and will aid in elucidating the constitution of regulons. There is still a great deal of skepticism about such *in silico* methods. The results obtained using motif-detection algorithms depend to a large extent on selecting the right parameter settings. This usually requires extensive parameter fine-tuning and user experience and could be discouraging. The large influence that using the appropriate background model has on the predictive capacity of the algorithm urged us to extend our Motif Sampler with background models of all

the sequenced prokaryotes. We believe that continuously updating and adapting motif-detection methods will enhance their user-friendliness and eventually alleviate the reluctance of researchers to use these *in silico* methods.

Acknowledgements

This work is partially supported by the IWT (projects STWW-00162, STWW-Genprom and GBOU-SQUAD); the Research Council KULeuven (projects GOA Mefisto-666 and IDO genetic networks); the FWO (projects G.0115.01 and G.0413.03); and IUAP V-22 (2002–2006). K.M. and S.D.K. are supported by the Fund for Scientific research (FWO-Vlaanderen) and G.T. by the IWT.

References

1 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371
2 Moreau, Y. *et al.* (2002) Functional bioinformatics of microarray data: from expression to regulation. *IEEE Proceedings* 11, 1722–1743
3 Manson, M.A. and Church, G.M. (2000) Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucleic Acids Res.* 28, 4523–4530
4 McCue, L.A. *et al.* (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* 12, 1523–1532
5 Lawrence, C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214
6 Liu, X. *et al.* (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 6, 127–138.http://psb.stanford.edu/psb-online/
7 Bailey, T.L. and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, 21–29
8 Thijs, G. *et al.* (2002) A Gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* 9, 447–464
9 Workman, C.T. and Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.* 5, 467–478.http://psb.stanford.edu/psb-online/
10 McCue, L. *et al.* (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* 29, 774–782
11 McGuire, A.M. *et al.* (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10, 744–757
12 Thijs, G. *et al.* (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17, 1113–1122
13 Thijs, G. *et al.* (2002) INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics* 18, 331–332
14 Barrios, H. *et al.* (1999) Compilation and analysis of σ(54)-dependent promoter sequences. *Nucleic Acids Res.* 27, 4305–4313
15 Dombrecht, B. *et al.* (2002) Prediction and overview of the RpoN-regulon in closely related species of the Rhizobiales. *Genome Biol.*, research00761–researc007611
16 Reitzer, L. and Schneider, B.L. (2001) Metabolic context and possible physiological themes of σ(54)-dependent genes in *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* 65, 422–444
17 Benson, D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.* 30, 17–20