

regulates at least 82 genes in *M. tuberculosis* and some of these genes encode proteins involved in nitrogen metabolism during late stationary phase [8]. Interestingly, these SigJ-regulated genes have also been lost as pseudogenes in *M. leprae*. Both SigJ and SigH have higher rates of stop codon accumulation compared with the genes they regulate, as shown in Fig. 2. in the supplementary material online. These observations fit the proposed model very well. As the experimental methods used in this research detect only highly expressed genes, one cannot immediately obtain estimates for the number of genes expressed in low quantities. This is an important factor that must be considered to obtain a correct estimate of the number of genes regulated by these alternative sigma factors. Assuming that each set of alternative sigma factors regulates approximately the same number of genes, one could arrive at a conservative estimate that the alternative sigma factors could regulate ~480–820 genes in all.

Conclusion

In conclusion, I propose that the loss of a set of sigma factors could have been the triggering step for the accumulation of pseudogenes in the *M. leprae* genome. A further set of sigma factor inactivation events could have occurred at a later point in time, shutting down the expression of another set of genes and forcing the pathogen to adopt a more specialized environmental niche for survival. This set of genes would now start to accumulate mutations. If this scenario occurred, at this point in time the latter subset would have accumulated fewer mutations than the former set for proteins of similar length, suggesting that pseudogene accumulation could have been triggered by at

least two independent events by the loss of sets of sigma factors.

Acknowledgements

I would like to thank Drs Teichmann, Sankaran and Rogozin for reading the manuscript and the anonymous referees for their very useful comments. I am grateful to the Medical Research Council, Cambridge Commonwealth Trust and Trinity College, Cambridge for financial support.

References

- 1 Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011
- 2 Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544
- 3 Lewin, B. (1998) *Genes VI*, Oxford University Press
- 4 Missiakas, D. and Raina, S. (1998) The extracytoplasmic function sigma factors: role and regulation. *Mol. Microbiol.* 28, 1066–1069
- 5 Petrov, D.A. *et al.* (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287, 1060–1062
- 6 Eiglmeier, K. *et al.* (2001) The decaying genome of *Mycobacterium leprae*. *Lepr. Rev.* 72, 387–398
- 7 Manganelli, R. *et al.* (2002) Role of the extracytoplasmic-function sigma factor, SigH in *Mycobacterium tuberculosis* global gene expression. *Mol. Microbiol.* 45, 365–374
- 8 Hu, Y. and Coates, R.M. (2001) Increased levels of sigJ mRNA in late stationary phase cultures of *Mycobacterium tuberculosis* detected by DNA array hybridization. *FEMS. Microbiol. Lett.* 202, 59–65

Supplementary material

Supplementary material and the dataset used for this analysis is available at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/mlep_pseudo/

0966-842X/03/\$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0966-842X(02)00031-8

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

Kathleen Marchal¹, Gert Thijs¹, Sigrid De Keersmaecker², Pieter Monsieurs¹, Bart De Moor¹ and Jos Vanderleyden²

¹ESAT SISTA-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

²Centre 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.

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

<http://timi.trends.com>

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. Bioprosector [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 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

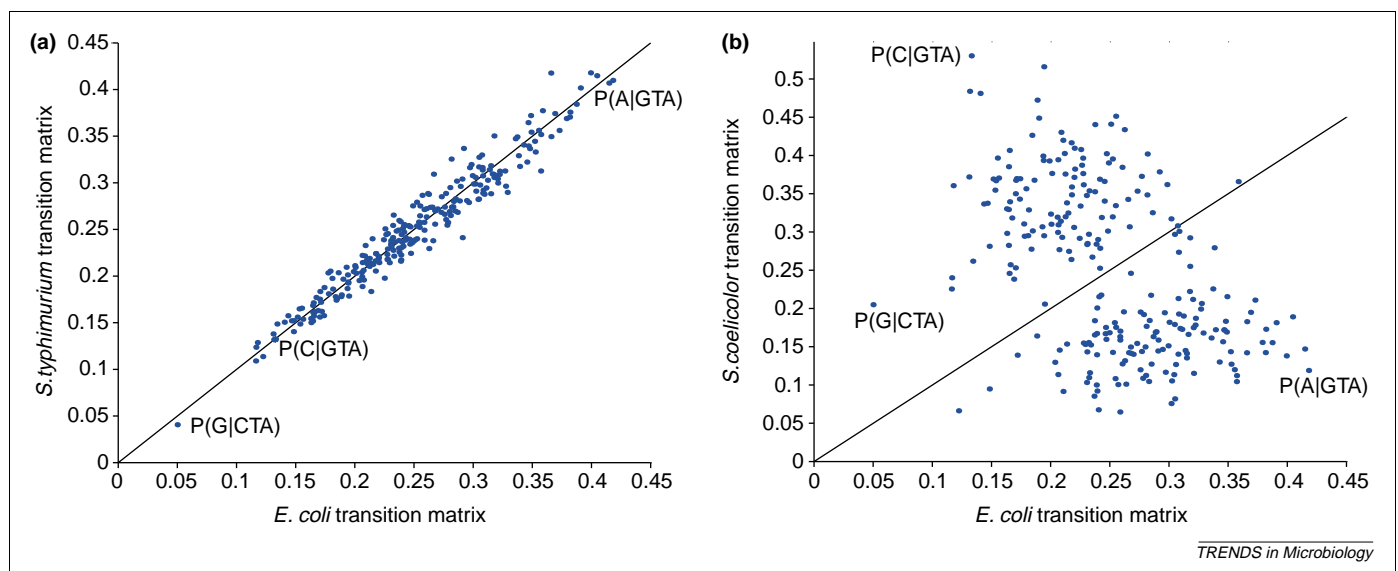


Fig. 1. Comparison of 3rd-order transition probabilities of two different genomes. Each dot corresponds to an entry in the transition matrix, which gives the probability of a nucleotide given the preceding trimer. (a) Plot of the transition probabilities of genomes with similar background composition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

^aThe functional annotation of each gene, derived from GenBank [17].

^bThe length of the intergenic region used for motif detection.

^cThe 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]).

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],

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}	
<i>E. coli</i> dataset										
Order 3, length 17										
<i>E. coli</i>	TGGCACrAywmnTGCAT	167.48	167.48	1.0969	1.0969	1	13	37		
<i>L. monocytogenes</i>	TGGCACrAywmnTGCAT	179.44	179.44	1.0969	1.0969	1	13	36		
<i>P. aeruginosa</i>	wnwwwAnnnmATwATw	151.02	–	0.5118	–	–	63	0		
<i>S. coelicolor</i>	wwwwnynkymwnnwnw	200.21	–	0.2846	–	–	132	0		
Order 0, length 17										
<i>E. coli</i>	TGGCACrAywmnTGCAT	177.17	177.17	1.0969	1.0969	1	13	23		
<i>L. monocytogenes</i>	TGGCACrAnwnnTGCwT	195.17	195.17	1.076	1.076	1	14	57		
<i>P. aeruginosa</i>	wwwwnAKsrmAnnwww	159.87	–	0.4539	–	–	79	0		
<i>S. coelicolor</i>	wwwwnynnnmwnwww	193.54	–	0.2862	–	–	132	0		
Order 3, length 7										
<i>E. coli</i>	TGGCACr	121.37	121.37	1.5467	1.5467	1	17	23		
<i>L. monocytogenes</i>	TGGCACr	132.2	132.2	1.5015	1.5015	1	19	53		
<i>P. aeruginosa</i>	TwmTTAA	122.71	–	1.1472	–	–	47	0		
<i>S. coelicolor</i>	nwnwnAw	145.64	–	0.7037	–	–	138	0		
Order 0, length 7										
<i>E. coli</i>	TGGCACr	124.16	124.16	1.5467	1.5467	1	17	21		
<i>L. monocytogenes</i>	TGGCACr	131.43	131.43	1.5467	1.5467	1	17	52		
<i>P. aeruginosa</i>	wTAAmAr	122.36	–	1.0184	–	–	63	0		
<i>S. coelicolor</i>	ATwwTnA	139.4	–	0.8052	–	–	124	0		
<i>P. aeruginosa</i> dataset										
Order 3, length 17										
<i>E. coli</i>	yCGsGsCsknCsnng	152.86	–	0.5681	–	–	83	0		
<i>L. monocytogenes</i>	CGsnGmnsnnnsCs	194.07	–	0.3646	–	–	192	0		
<i>P. aeruginosa</i>	nTGGCACGsnwTTGCT	142.92	142.92	1.0869	1.0869	1	11	37		
<i>S. coelicolor</i>	nwwwkrnwryrynAwnn	178.09	–	0.3791	–	–	71	0		
Order 0, length 17										
<i>E. coli</i>	ssnnGsCnnnsnCssn	167.62	–	0.4947	–	–	121	0		
<i>L. monocytogenes</i>	ssnnsnsnnnsns	204.6	–	0.3313	–	–	211	0		
<i>P. aeruginosa</i>	TTsywnynyTTGnnn	134.1	124.65	0.7565	1.2322	4	17 (8)	14		
<i>S. coelicolor</i>	nwwwTnnwnnTwnTnr	158.52	–	0.42321	–	–	60	0		
Order 3, length 7										
<i>E. coli</i>	CGCGsCk	126.41	–	1.3551	–	–	47	0		
<i>L. monocytogenes</i>	sCsnGsC	153.9	–	1.1023	–	–	125	0		
<i>P. aeruginosa</i>	wTTGGCA	111.43	111.43	1.4142	1.4142	1	16	15		
<i>S. coelicolor</i>	nTkmwAw	121.08	–	0.0808	–	–	66	0		
Order 0, length 7										
<i>E. coli</i>	ssCGGcs	140.69	–	1.2654	–	–	82	0		
<i>L. monocytogenes</i>	sCsCGsC	162.26	–	1.0372	–	–	154	0		
<i>P. aeruginosa</i>	TTTTnck	110.66	109.06	1.2836	1.416	2	23 (18)	4		
<i>S. coelicolor</i>	WmTwwTT	118.46	–	0.8946	–	–	56	0		

^aThe 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.

^bThe 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.

^cExpresses the number of occurrences 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).

^dExpresses the number of times the σ^{54} motif was recorded on 100 runs of the algorithm.

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 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 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 $\sigma(54)$ -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. research*0076.1–0076.11
- 16 Reitzer, L. and Schneider, B.L. (2001) Metabolic context and possible physiological themes of $\sigma(54)$ -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

0966-842X/03/\$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0966-842X(02)00030-6

Microbial Genomics

The building blocks of pathogenicity

Nicholas R. Thomson and Julian Parkhill

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

As the workhorse of bacterial molecular genetics, the publication of any new *Escherichia coli* genome sequence is always greeted with great interest. Welch *et al.* [1] have provided us with the complete sequence of another *E. coli* genotype, namely the uropathogenic *E. coli* (UPEC) strain CFT073. The previously sequenced genomes were that of the relatively benign laboratory strain K12 [2], and two independent publications of enterohaemorrhagic *E. coli* (EHEC) 0157:H7 strains EDL933 and Sakai [3,4]. UPEC are a diverse group of extraintestinal pathogens, able to live harmlessly as commensals within the gut as well as being able to colonize a variety of other niches within the human body including the perineum, bladder, urethra and kidneys.

A comparison of the genome sequences of UPEC and K12 echoed many of the same themes seen in the EHEC vs K12 comparison [3]. Essentially, UPEC and K12 share a common 3.9 Mb core of DNA, which is highly conserved, both in gene order and percentage identity (~98% sequence identity), and encodes mostly housekeeping functions. This core or backbone sequence is punctuated by many discrete insertions (UPEC-specific islands; UIs) totalling 1.3 Mb (25% of the total genomic DNA). Analysis of the codon usage of the UIs showed that not only was it atypical, with respect to the backbone sequence, but there was also a preponderance of rare codons. Together, these observations were taken to suggest that UIs have been acquired from a foreign source. Moreover, many of the larger UIs were found to encode known or predicted virulence functions. Consequently, UIs have been attributed with giving UPEC the ability to colonize the urinary tract and cause an array of diseases, including cystitis, neonatal meningitis and pyelonephritis.

Welch *et al.* compared the coding sequences of all three of the sequenced *E. coli* strains and grouped the orthologous proteins. The results showed that the core gene set shared by K12, EHEC and UPEC genotypes comprises 2996 genes. The genes present in UPEC but absent from K12 numbered 1827; the same figure for EHEC was 1387 [3] and even K12 has 585 genes not present in the other two *E. coli* strains. It is clear from this



that the divergence of the pathogenic *E. coli* from K12 has taken an expansionist path. Interestingly, only 11% of the 1827 genes unique to UPEC were also found in EHEC. This illustrates the scale of the genetic diversity within these three genotypes and, assuming the DNA has been acquired laterally, the level of the genetic interchange possible in an environment like the digestive tract.

The virulence-related functions are mainly encoded within the UIs and include 12 distinct fimbrial systems, such as the two *pap* operons, which are known to be uropathogen specific. Several other fimbrial systems identified in UPEC, such as the *yad* chaperone-usher system, are also present in the other two sequenced genotypes. However, even these ubiquitous fimbrial systems display a level of sequence variation, which suggests that they interact with different target sites. UIs were also found to carry seven autotransporters, a novel RTX-family toxin and five *fimE* and *fimB* recombinase systems, all of which are associated with aspects of host interaction or, in the case of the recombinases, phase switching.

As previously mentioned, the core UPEC genes are highly conserved and, although not generally associated with virulence, are thought to act as integration sites for the laterally acquired DNA. Consistent with this idea is the fact that 13 of the UIs identified within the genomic sequence were found to be integrated alongside tRNA

Corresponding author: Nicholas R. Thomson (microbes@sanger.ac.uk).