

Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint

Brian Grech^a, Stefan Maetschke^b, Sarah Mathews^a, Peter Timms^{a,*}

^a Institute of Health and Biomedical Innovation, Queensland University of Technology,
Corner Musk Avenue and Blamey Street, Kelvin Grove, Brisbane, Queensland, 4059, Australia

^b Faculty of Information Technology, Queensland University of Technology, 126 Margaret Street, Brisbane, Queensland, 4001, Australia

Received 25 June 2007; accepted 22 August 2007

Available online 18 October 2007

Abstract

Currently, there is a lack of phylogenetic footprinting programmes that can take advantage of multiple whole genome sequences of different species within the same bacterial genus. Therefore, we have developed and tested a position weight matrix-based programme called *Footy*, that performs genome-wide analysis of bacterial genomes for promoters that phylogenetically footprint. When *Footy* was used to analyse the non-coding regions upstream of genes from three chlamydial species for promoters that phylogenetically footprint, it predicted a total of 42 promoters, of which 41 were new. Ten of the 41 new promoters predicted by *Footy* were biologically assayed in *Chlamydia trachomatis* by mapping the 5' end of the transcripts for the associated genes. The primer extension assay validated seven of the 10 promoters. When *Footy* was compared to two other accepted methods for genome-wide prediction of promoters in bacteria (the standard PWM method and MITRA), *Footy* performed equally as well or better than these programmes. This paper, therefore, shows the value of a bioinformatics programme able to perform genome-wide analysis of bacteria for promoters that phylogenetically footprint.

© 2007 Elsevier Masson SAS. All rights reserved.

Keywords: Bioinformatics; *Chlamydia*; Phylogenetic footprinting; Promoter

1. Introduction

The post-genomic era has generated an interest in developing theoretical approaches to discover as much information as possible from the available genome sequences [16]. One area of intense research is modelling and predicting bacterial promoters. However, the structure of bacterial promoters makes it difficult to devise a general prediction algorithm [8,10,11]. Most of the currently available programmes for the prediction of bacterial promoters exhibit poor specificity, generating many false-positive predictions. One approach to filter out false-positives predicted by the current methods is phylogenetic footprinting [8].

Phylogenetic footprinting is a computational method for predicting homologous promoters in the equivalent non-

coding regions (NCRs) upstream of a gene family (or genes with a common function), from evolutionarily related species [6]. (NCRs upstream of genes will be referred to as *upstream regions* in this paper.) Phylogenetic footprinting predicts promoters by assuming that: (1) the upstream regions of homologous genes from different species are regulated by homologous promoters, and (2) spacers between promoters within the equivalent upstream regions are free from evolutionary constraints. Therefore, substitution of another base within the spacer can be accepted at any position, whereas the homologous promoters can only accept certain substitutions, as they must be recognizable to the cognate σ factor or transcription factor.

The increasing number of complete bacterial genome sequences now allows genome-wide analysis for promoters that phylogenetically footprint. This is because it is possible to find a dataset of equivalent upstream regions from two or more bacterial species separated by an appropriate evolutionary

* Corresponding author. Tel.: +61 7 3138 6199; fax: +61 7 3138 6030.

E-mail address: p.timms@qut.edu.au (P. Timms).

distance for phylogenetic footprinting of promoters [6,7,34]. The appropriate evolutionary distance is observed within a particular evolutionary time frame. For example, a dataset of equivalent upstream regions from too *closely related species* will produce a high rate of false-positives, when phylogenetically footprinting promoters, since the spacers surrounding the promoters have not had enough evolutionary time to mutate, so that the conservation of the spacers across the different species is poorer when compared to the conservation of the promoters across the same species. At evolutionary distances that are too great, only *well conserved* promoters, upstream of well conserved genes can be phylogenetically footprinted. The appropriate evolutionary distance is when there is an observable difference between the conservation of the spacers and the conservation of the promoters across the different species [6,34].

Bacteria of the genus *Chlamydia* are ideal organisms for finding a species set at an appropriate evolutionary distance for phylogenetically footprinting promoters. This is because chlamydiae are phylogenetically isolated, which has resulted in a high level of conservation of genes and gene order between the species [1,22] and *Chlamydia* is one of the most sequenced organisms with 10 genome sequences available for six species [1,4,9,13,21,22,26,27,32]. Consequently, many of the promoters would be expected to be well conserved across the different chlamydial species and the probability of finding these well conserved promoters would be high.

Chlamydial σ^{66} promoters are probably the best choice of promoters to be identified by phylogenetic footprinting. This is because σ^{66} promoters have a greater likelihood to be found upstream of the majority of chlamydial operons (since, the σ^{66} factor (RpoD) of *Chlamydia*, is the principal σ factor [14,18].) Promoter mutagenesis and in vitro transcription assays have shown that σ^{66} have the greatest affinity for two motifs that are identical to the -35 and -10 hexamers of the *Escherichia coli* σ^{70} consensus sequence and are, therefore, σ^{70} -like [30].

The accepted method for predicting *E. coli* σ^{70} promoters is to use a pair of position weight matrices (PWMs) to predict the -35 and -10 hexamers [28]. A PWM is a two-dimensional array of values representing the information content (IC) of a motif. The IC is a measure of the bit rate, i.e. bits per base. However, a phylogenetic footprinting algorithm that uses PWMs to analyse the upstream regions of multiple whole bacterial genomes for promoters that phylogenetically footprint has not been published.

The work presented in this paper has developed and tested a bioinformatic programme that can perform genome-wide analysis of bacteria for promoters that phylogenetically footprint. This programme is called *Footy* and is based on the standard PWM method, with an extension that can analyse multiple bacterial genomes for phylogenetically conserved promoters. When *Footy* was applied to the genomes of *Chlamydia trachomatis*, *Chlamydia pneumoniae* and *Chlamydia caviae*, 42 σ^{66} promoters were predicted, of which 41 were new.

2. Materials and methods

2.1. Sequence data

The plus strand of whole genomes of *C. trachomatis* serovar D, *C. pneumoniae* strain AR39, *C. caviae* biovar GPIC and *Chlamydia muridarum* biovar MoPn, and the corresponding annotation table were downloaded from GenBank (<http://www.ncbi.nlm.nih.gov>; accession nos.: AE001273, AE002160, AE002161 and AE015925, respectively). The coordinates used for each predicted gene and *structural RNAs* (rRNAs and tRNAs) were based on the annotated start and stop positions [21,22,27]. Two datasets of regions were generated from the four genomes: (1) upstream regions and (2) downstream regions (or NCRs between convergently transcribed genes). The 3' ends of the upstream regions were reduced by 10 bp from the start of the associate gene. The maximum length of each upstream region was limited to 390 bp and the minimum length was restricted to 35 bp. The maximum length of each downstream region was not limited, but the minimum length was restricted to 35 bp.

2.2. Identification of equivalent upstream and downstream regions

A table of predicted homologous genes (TOPHG) was constructed for *C. trachomatis*, *C. muridarum*, *C. pneumoniae* and *C. caviae* using prototype tables of homologous genes calculated by “TIGR Comprehensive Microbial Resource Total Protein Hit” search engine (<http://www.tigr.org/tigr-scripts/CMR2/>). The parameters used for BLAST analysis were as follows: similarity $\geq 40.0\%$, identity $\geq 10.0\%$ and P -value ≤ 0.05 . For duplicate entries with the same gene names, the set of homologous genes with the lowest P -value was selected. The sets of structural RNAs were identified by comparing the location of the genes downstream of the structural RNAs in *C. trachomatis* with their homologs in *C. muridarum*, *C. pneumoniae* and *C. caviae*. Candidate upstream and downstream regions were determined to be equivalent if their associated genes were predicted to be homologous, using the above parameters.

2.3. Footy

A flow chart of the *Footy* algorithm is shown in Fig. 1. The promoter model consisting of two PWMs and variable spacer was calculated. The PWMs were derived from an alignment of the -35 and -10 hexamers of 300 *E. coli* σ^{70} promoters taken from Lissner and Margalit [17]. The weights of the PWMs were calculated using equations (equations S1, S2 and S3, supplementary material) based on the equations developed by Stormo and Hartzell [29]. The first stage of *Footy* scans the upstream regions of the chlamydial genomes for patterns that are similar to the model. The first stage is the same as the standard PWM method. The second stage of *Footy* phylogenetically footprints promoters with homologous promoters

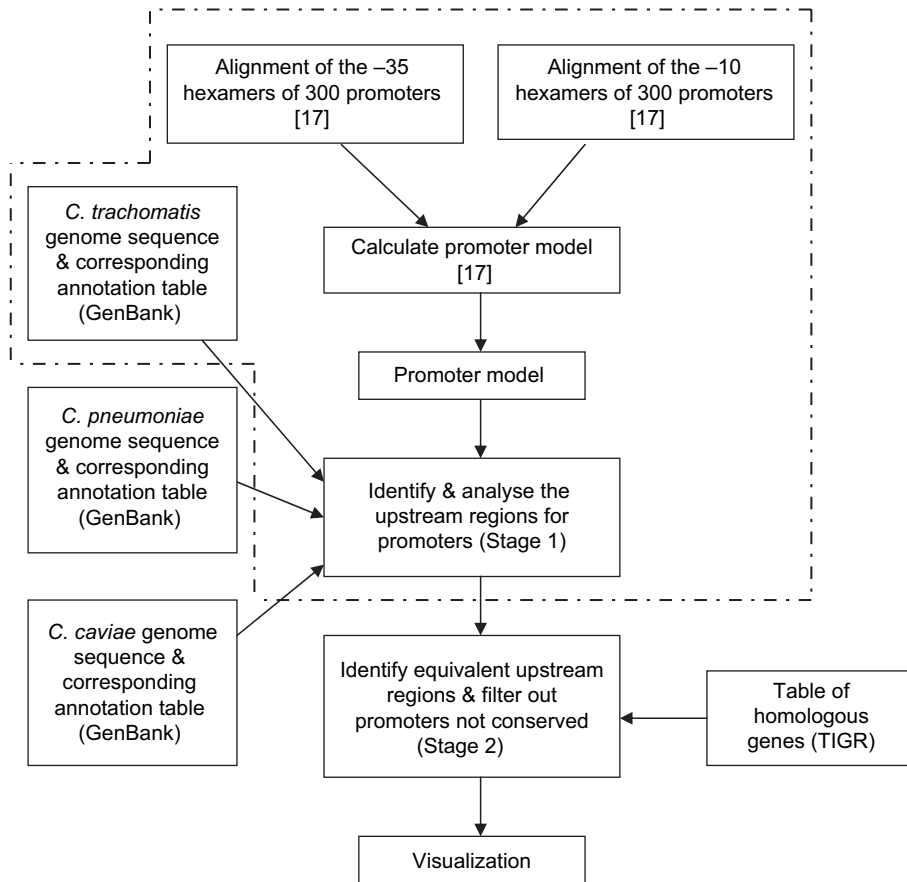


Fig. 1. Schematic representation of the steps of Footy. The section within the dot-dashed line shows the steps of the standard PWM method.

in the other chlamydial species. Footy with instruction on usage is available at <http://ereseach.fit.qut.edu.au/Footy/>.

2.4. Reduction of false-positives

The second stage of Footy eliminates many of the promoters predicted that did not phylogenetically footprint by aligning the predicted -35 and -10 hexamers in *C. trachomatis* with the predicted hexamers for the homologous genes (where available based on the TOPHG) in the other chlamydial species (Fig. 2). To do this Footy performed un-gapped pair-wise alignments (PWAs) of the predicted hexamers. These regions were aligned at the first base of each predicted hexamer. The predicted hexamers were reported as conserved if the number of mismatches between the reference species and the other species were equal to or less than the pre-set mismatch threshold. If all of the PWAs reported conserved hexamers equal to or less than the mismatch threshold, then the hexamers were reported as well conserved (Fig. 2A). Once this process was completed, the next set of equivalent upstream regions were analysed (Fig. 2B).

To further filter out false-positives, the multiple sequence alignments (MSAs) calculated by Footy were inspected to decide which, if any, of the predicted well conserved promoters could be eliminated. The position of each predicted promoter with respect to the start site of the associated gene was

compared between species. If the distances varied more than 200 bp between different species, the promoters were eliminated. If multiple promoters were predicted in the same upstream regions, the highest bit scoring and lowest mismatch MSA of conserved promoters was selected.

2.5. Validation of a subset of predicted promoters

To validate the predicted promoters, a subset were chosen to have the 5' end of the RNA of the associated genes mapped by primer extension in *C. trachomatis* serovar L2/434/Bu (Table S1) [19]. The predicted promoter was considered to be correct if the spacer between the -10 hexamer and the mapped 5' end of the RNA was from 4 to 12 bp [12].

3. Results

3.1. Footy predicted 42 promoters that were phylogenetically conserved in Chlamydia

To determine the number of species chosen at an appropriate evolutionary distance, IC threshold and mismatch threshold, a dataset of equivalent downstream regions of *Chlamydia* was analysed for false-positives using a σ^{70} promoter model. The analysis of the downstream regions revealed a lack of false-positives (data not shown); therefore, the downstream

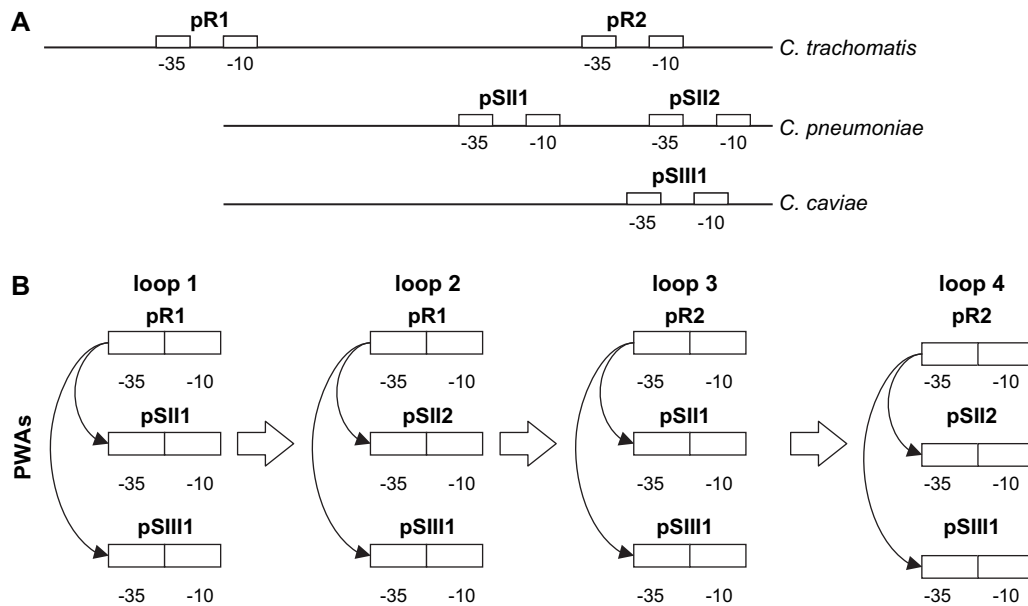


Fig. 2. Representation of how the filtering algorithm of Footy performs PWAs of all the predicted promoters within the equivalent upstream regions. (A) Diagrammatic representation of promoters predicted within a dataset of three equivalent upstream regions. Two boxes (–35 and –10 hexamers) represent each predicted promoter, which are numbered in order of their prediction (e.g. pR1 and pR2), within each upstream region. (B) Illustration of the order in which the ungapped PWAs were performed. Hexamers predicted in part A were transferred to part B and pasted together. The line arrows represent the order in which the PWAs occurred within each loop. The boxed arrows represent the order in which different combinations of PWAs occurred between loops.

regions could not be used to provide an appropriate negative control. While performing this analysis it became clear that a 16–18 bp spacer between the two hexamers, rather than a 15–21 bp spacer, which is the allowable spacer length for *E. coli* σ^{70} promoters [17], substantially reduced the number of false-positives predicted (data not shown).

The number of species chosen, IC threshold and mismatch threshold were set so that the maximum number of promoters were predicted and no false-positives were detected in the upstream regions of a dataset of 15 positive controls. The 15 positive controls chosen were the *C. trachomatis* 16S rDNA P1, CT602, *hctA*, *infA*, *ltuA*, *ltuB*, *omcA*, *ompA* P1 and P2, *pkn5*, *rpoD*, *secA*, *sctU*, *srp* and tRNAThr2 (B.J. Grech unpublished data), [20]. These σ^{66} promoters were determined to be a suitable set of positive control promoters for setting parameters and testing the performance of Footy, because they were located in NCRs on the *C. trachomatis* chromosome. The analysis revealed that using the species of *C. trachomatis*, *C. pneumoniae* and *C. caviae*, with an IC threshold of 3.0 bits (or 1.5 bits per hexamer) and a mismatch threshold of two resulted in the maximum number of promoters predicted with none of the 15 positive controls reported as false-positives. Table 1 shows the number of false-positives for different combination of species, IC thresholds and mismatch thresholds.

Using these parameters and after applying the rules discussed in Section 2, Footy predicted 42 promoters that were conserved in the dataset of 305 equivalent upstream regions extracted from the three chlamydiae. One of the 15 positive control promoters, the promoter of *infA*, was predicted correctly by Footy. Footy did, however, predict three new promoters in the upstream regions of *euo*, *groES* and *rpsA*,

with homologs that have been biologically confirmed in chlamydial species not analysed in this study. The –35 and –10 hexamers of the predicted promoters of *groES* and *rpsA* were 100% conserved with the –35 and –10 hexamers of the biological confirmed promoters of *groES* and *rpsA* of *C. muridarum* [31]. The predicted promoter of *euo* was 4 bp upstream of the nucleotide in *C. trachomatis* that corresponds to the 5' end of the *euo* P1 transcript, mapped in *C. psittaci* 6BC by Wichlan and Hatch [35] (Table 2).

An analysis was conducted to determine why 14 of the 15 positive controls were missed by Footy. The *C. trachomatis*, *C. pneumoniae* and *C. caviae* genomes were visually inspected for patterns similar to the 14 false-negatives. Analysis of the equivalent regions in *C. pneumoniae* and *C. caviae* for patterns similar to the 14 false-negatives, identified patterns with no more than two mismatches from the promoters of *C. trachomatis* CT602, *hctA*, *omcA*, *rpoD* and *sctU*. The promoters CT602, *hctA*, *rpoD* and *sctU* were missed because they were below the (3.0 bit) IC threshold and the promoter for *omcA* was missed because the analogous pattern in the equivalent upstream region of *C. pneumoniae* was located within the open reading frames (ORFs) (Table 3).

The total number of mismatches that the –35 and –10 hexamers of each of the 42 promoters had from the σ^{70} consensus sequence (TTGACA and TATAAT, for –35 and –10, respectively) were determined. The number of mismatches ranged from zero to five out of a possible 12, with a statistical mode of four mismatches. Fifteen (35%) of the promoters had four mismatches, 34 (80%) of the promoters had three to five mismatches and 41 (98%) of the promoters had one to six mismatches from the σ^{70} consensus sequence.

Table 1
Results of the analysis of the upstream regions of different combination of chlamydial species for positive control promoters

1 IC threshold (bits)		7.5												7.0																							
2 Species		D & MoPn				D & GPIC				D & AR39				D & MoPn				D & GPIC				D & AR39				D, GPIC & AR39				D, MoPn, GPIC & AR39							
3 Mismatches		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
4 False positives																														✓	✓						
5 Number of positive controls predicted		3 3								3 3								3 3				4 4				4				4 4							
6 Number of promoters predicted		30 31								35 44								14 14				18 24				29				16 19							

IC threshold (bits)		6.5												3.0																											
Species		D & MoPn				D & GPIC				D & AR39				D, GPIC & AR39				D, MoPn, GPIC & AR39				D & MoPn				D & GPIC				D & AR39				D, GPIC & AR39				D, MoPn, GPIC & AR39			
Mismatches		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
False positives		✓	✓					✓	✓					✓	✓					✓	✓																				
Number of positive controls predicted																		3 4				4 6				3 5 5				5 5											
Number of promoters predicted																		33 65				24 53				19 35 44				23 46											

1 IC threshold (bits)		2.5																			
2 Species		D & MoPn				D & GPIC				D & AR39				D, GPIC & AR39				D, MoPn, GPIC & AR39			
3 Mismatches		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
4 False positives						✓	✓					✓	✓					✓	✓		
5 Number of positive controls predicted																					
6 Number of promoters predicted																					

Row 1: the sum of the IC threshold of the -35 and -10 hexamers used for the analysis of the dataset upstream regions of chlamydiae. Row 2: species analysed; species abbreviations are as follows: AR39, *C. pneumoniae*; D, *C. trachomatis*; GPIC, *C. caviae*; and MoPn, *C. muridarum*. Row 3: the mismatch threshold used for the analysis. Row 4: indicates the IC threshold (Row 1) species combination (Row 2), and mismatch threshold (Row 3) where the first false-positive was reported. Row 5: the number of positive controls predicted. Row 6: the number of promoters predicted. (Data shown in Rows 5 and 6 are at an IC threshold 0.5 bits more than the IC threshold at which the first false-positive was reported.)

3.2. Transcription start site mapping confirms an additional seven promoters predicted by Footy

Since only one of the 42 predicted promoters predicted by Footy had been biologically confirmed, more experimental data was needed to assess the performance of Footy. Therefore, 10 of the 42 promoters predicted by Footy, the promoters of *tyrS*, *gcp1*, *clpC*, *rs12*, *CT547*, *sctJ*, *exbB*, *snf*, *greA* and *efp2*, were chosen for primer extension. The 10 genes were selected since they are highly expressed at 24 h post infection by microarray analysis [2], thus ensuring gene specific RNA would be isolated.

The 5' end for *tyrS*, *clpC*, *rs12*, *sctJ*, *exbB*, *snf* and *elp2* transcripts correctly mapped to the promoters predicted by Footy (Table 2 and Fig. S1, supplementary material), hence confirming an additional seven promoters. The promoters of the remaining three genes (*gcp1*, *CT547* and *greA*) were unable to be confirmed, because the 5' end of RNA could not be mapped or was mapped elsewhere (data not shown). Since

C. trachomatis was grown in HEp2 monolayers there was the possibility of non-specific binding of the gene-specific primers to HEp2 RNA. Therefore, primer extension was also performed on RNA extracted from uninfected HEp2 cell lines and all results were negative (data not shown).

The predicted -35 and -10 hexamers of the seven newly confirmed promoters (*tyrS*, *clpC*, *rs12*, *sctJ*, *exbB*, *snf* and *elp2*) were analysed and compared to *E. coli* σ^{70} promoters. The bit scores for both hexamers ranged from 4.4 to 9.6 bits, out of a range of 3.0 to 14.4 bits and the number of mismatches for both hexamers across the three chlamydiae ranged from zero to the maximum mismatch threshold of two. The promoters of *efp2*, *rs12* and *sctJ* had nine nucleotides; the promoters of *snf* and *tyrS* had eight nucleotides and the promoters of *clpC* and *exbB* had seven nucleotides identical to the σ^{70} consensus sequence. Since σ^{70} promoters can have as few as five nucleotides identical to the consensus sequence [17], the seven newly identified promoters are σ^{70} -like and classified as σ^{66} promoters.

Table 2
 σ^{66} Promoters predicted by Footy in the upstream regions of *C. trachomatis* D

Name	Location	Sequence	-35 hexamer	-10 hexamer	IC (bits)	Mismatches	Putative product
CT017	18418	ATCGAACAGTCCGAG TTGACTTTTTCTCTT	TTGACTTTTTCTCTT	AAGTCAATA ATAAT TCCCTCTCTA	9.0	1	hypothetical protei n
CT043	48617	TCCGGCGCTCTAAT CTTACTAACAAACCT	CTTACTAACAAACCT	GCTTATGCTAGG TTTAA AAAAAAC	5.5	2	hypothetical protei n
CT062-tyrS	71848	CTGCTATCGCTTGCC TTGCTATAAAAAGAAC	TTGCTATAAAAAGAAC	AGGATAGA TAAGAT GTGGCAGAT	7.5	2	thro syl tRNA synthetase
CT066	799156	AAAGATAAACACTAAT TGATT TTTTATTTC	TGATT TTTTATTTC	ACTGAACAT TAAT CGAAAAAAC	5.9	2	hypothetical protei n
CT098-rs1	115695	AGTCAAGGGAAAT CTTGCC TTTTTAAGG	CTTGCC TTTTTAAGG	TGAATAT TTACACTACTCT ctcttTG	5.9	0	30S ribosomal protein S1
CT111-groES	128417	CAACTGCTAAAC CCAGTTGC AAAAAGCGAG	CCAGTTGC AAAAAGCGAG	GACT TTGCTATCGT TCTTCCtCTG	7.5	1	10 kD chaperonin (hea t shock protein GroES)
tRNAAla_1	202535	TTTGATAAT CTTTCTT TTGGCTTAAATCGC	CTTTCTT TTGGCTTAAATCGC	TCTGGAT TAAGAT GGCGCTTTGT	5.1	2	tRNA Ala
CT197-gcpl	221387	CTTGGATTAA CGCC TTGCTTGATTAACAA	CGCC TTGCTTGATTAACAA	TCTCATGAT TACGAT CCTCTCCTTC	5.8	2	O-sialogl ycoprotein endopeptidase
CT265-accA	297412	TAAGAGAAAT TAATA TTGTTGCGTGAA	TAATA TTGTTGCGTGAA	AAGGTCAT TATAAT CAATAGTTG	8.7	2	acet yl-CoA carb oxylase transf erase (α subunit)
CT266	298721	TCAGCGTAAGCAAG CTTGACT CTAAATTTTC	CTTGACT CTAAATTTTC	CTCAAGAT TATTT TTGCCATTTGG	8.7	2	hypothetical protei n
CT267-ihfA	299179	AAGATAAAAAG CTCTT GAATCCAAAGGA	CTCTT GAATCCAAAGGA	TGAATGCAT TATTAT ACGCATATAT	8.6	1	hist one-like DN A binding proteins, IH FA, IHFB or DBH
CT269-murE	301529	GTTAGTCGCAAA AGCTTG CAACGAATAT	AGCTTG CAACGAATAT	GTGTATAG TAACT ATTTGAGAAA	11.5	2	UDP-N-acetylmu ram oylalanylglutamy l DAP Ligase
CT273	305141	TCCCTCTACCA TACTTGACT TTTTTCCT	TACTTGACT TTTTTCCT	CCCCCGAT TATGAT TGAGATTGG	11.3	1	Chlamydia-specific hyp othetical protei n (basic)
CT286-clpC	317907	TCCCTTTACGAAA AGTTGC ATCATTATCAT	AGTTGC ATCATTATCAT	AAATCTGT ATAATG CTTGAAAAAT	4.4	0	CLP protease e ATP-binding subunit (CLP A/CLPB/CLPC)
CT297-rnc	330617	AACCTGAAAG TACTATAG ACTTTAAGATT	TACTATAG ACTTTAAGATT	TTTCGGCT TATAAA ACCCCGATTG	5.4	0	ribonuclease III
CT323-infA	363851	TTTTTGACAA GTCTT GCATTTTCTGT	GTCTT GCATTTTCTGT	TTAGTCGAT TATAAT CGCTCTCTCG	14.4	1	translation initiation factor IF-1
CT342-rs21	391021	CAACTTAAGTAT CTCTTGAAG CCTAAATAAA	CTCTTGAAG CCTAAATAAA	AAAGTGT TACAAT CCCCGGTCTC	9.4	0	30S ribosomal protei n S21
CT393-proS	447959	AAAAATC ACAGAGAT TGATCTAGAAACAC	ACAGAGAT TGATCTAGAAACAC	TCCTATG TAAGAT GCTCTCCAC	7.2	0	Prolyl tRNA synthetase
CT398	156056	TTAAA CAAAACG TGCTTACTCTTGCGAGA	CAAAACG TGCTTACTCTTGCGAGA	AAAA TCGGTAACT TCGGCTTCG	7.6	1	conser ved hyp othetical protei n
CT439-rs12	508566	CCTAGAA ATAACCC CTTGCAACAAAGATAT	ATAACCC CTTGCAACAAAGATAT	TCTTAT CTATAT TCCCTGTTTG	9.6	1	30S ribosomal protei n S12
CT446-euo	517160	TTTTTAA CAAAACCG CTTGATTAATAAGTTT	CAAAACCG CTTGATTAATAAGTTT	TTTT GGGAAAAT gttaccCTTCT	5.8	0	CHLPS Euo protei n
CT455-murA	530787	TATTTTAT TTGTTT TAAAAACAAACAAAT	TTGTTT TAAAAACAAACAAAT	GTCTCT TTGTTAATA AGATGTTTT	4.0	2	UDP-N-Acetylglucosamine Transferase
CT475-pheT	548337	CTCCCT CAAAAT ACTTGCTACTATACAG	CAAAAT ACTTGCTACTATACAG	CCACT TCGTA AAATCTACAAAAA	9.1	1	phenylalanyl tRNA synthetase beta
CT496-pgsA1	574775	AGCTAA CTCTCTG CTTTGGAGTGT	CTCTCTG CTTTGGAGTGT	CTATG TTTCA TAATATGTGTCTATT	6.9	2	CDP-dia cygly cero l-gly cero l-3-phosphate 3-phosphati dyltransferase
CT528-r13	596666	GCTTAGCT TTTCTTAT TGTAATAATCTGTT	TTTCTTAT TGTAATAATCTGTT	TCCT TTGATAAT CTGTCCCTTTAA	5.3	1	L3 Ribosomal protei n
CT546	617372	AACAAAA AAATTTAT TGCGCATTCGCTCT	AAATTTAT TGCGCATTCGCTCT	TTAT TTTATAA ATAATAAAAAAG	4.3	2	Chlamydia-specific hyp othetical protei n
CT547	617442	AAAGCT ATAAGGAT TGCAAAATCTCTTT	ATAAGGAT TGCAAAATCTCTTT	TCCT TTTTATGAT GAGCCTTTGT	12.9	0	Chlamydia-specific hyp othetical protei n
CT559-actJ	631398	AAAATAT TTCCCGA TTGGCACTAATCTCC	TTCCCGA TTGGCACTAATCTCC	CCAT CTGCTATGGT GAGTCaAAAG	8.8	0	flagellar M-ring protei n (YopJ translocation protei n)
CT596-exbB	676817	ATACCA AAAAGGAT CTGGTCTATACAAG	AAAAGGAT CTGGTCTATACAAG	AAAT TTGT TAGGATCGTCTAGGAA	5.6	1	pol ysaccharide transporter
CT619	701690	TTATA AAAAACAAC ATAGAAAAAACTTTTT	AAAAACAAC ATAGAAAAAACTTTTT	TTAA ATAAGAAA ATAAAAAACATAA	4.2	0	hypothetical protei n
CT626-rs4	714150	AATCTAG GAATCCG TTGTAGAAAAATGGA	GAATCCG TTGTAGAAAAATGGA	AA TAGA ACTAGAACTGCTTTTGT	8.0	1	30S ribosomal protei n S4
CT636-greA	723227	TTATA AAAAACAAC ATAGAAAAAACTTTTT	AAAAACAAC ATAGAAAAAACTTTTT	TTAA ATAAGAAA ATAAAAAACATAA	7.8	0	transcription elongation factor (Gre A)
CT646	741250	TAATTA AGTTTTCT TGAAAAAGATGTTT	AGTTTTCT TGAAAAAGATGTTT	TTAT TTTTTAAA ATGAGCGCTCTT	10.9	0	Chlamydia-specific hyp othetical protei n
CT681-ompA	780229	GT TTTTCTT CTCAACT TTACG AGAATAAGAA	TTTTCTT CTCAACT TTACG AGAATAAGAA	AA TTTTT TATGGCTCTCGAGCAT	7.1	2	Major outer membrane protei n
tRNAGly_2	787678	TTCT CTAAA AGAGAT TGC ATAAAAATCCCT	CTAAA AGAGAT TGC ATAAAAATCCCT	GCT CTCCG TACTATATCGGCTAC	5.5	2	tRNA Gly
CT706-clpP2	813229	ATCGC AGGAAAC CGCTTGACCCAAAGAGACA	AGGAAAC CGCTTGACCCAAAGAGACA	CTTAA ACTAG AATTCATCATTTTT	11.4	0	ATP-dependent ClpP endopeptidase subunit
CT708-snf	814796	GGG CAAAATTT CTAGTATTAGCGGAAG	CAAAATTT CTAGTATTAGCGGAAG	TAAA AGGTACA AGTAAACAGATCT	5.2	0	proba ble helic ase
CT752-efp2	884446	TTCCG CACATTT CTGGACAAGCTTAGAA	CACATTT CTGGACAAGCTTAGAA	GAGA ACGATA AGCATAGATGGAGAA	8.1	0	Elongation factor P (EF-P)
CT768	901360	GATCC ATAA ACCG TTG CGATAATGCAT	ATAA ACCG TTG CGATAATGCAT	TGCC AGCAAA CTTTGACTACCA	7.9	1	hypothetical protei n
CT769-jbeB	903504	CT TTAG AAAAAG AGCTCG ACCTTATCTTAGA	TTAG AAAAAG AGCTCG ACCTTATCTTAGA	TT AGCGG TATCTCGAGGCAGTT	6.8	1	iojap s upefamily y ortholog
CT827-nrdA	974155	TATGCTAT TTTCAA TTG CAG GAACGTTG	TTTCAA TTG CAG GAACGTTG	CTAG CTTCTAT ATATGGTATACAA	4.6	1	ribonucleoside diphosphate reductase alpha chain
CT837	984553	TATAA ATAA ATAT TTG AAAAGCTAATTCAT	ATAA ATAT TTG AAAAGCTAATTCAT	TTATA AAAA TA AACT TAAGAACAAT	9.6	2	hypothetical protei n

Promoters were identified by the name of the associated gene. The location of promoters refers to the site on the *C. trachomatis* D genome of the first base of the predicted -35 hexamer. The bold uppercase nucleotides represent the predicted -35 and -10 hexamers and the bold lower case nucleotides represent the mapped TSSs (if available). Promoters aligned at the predicted -35 and -10 hexamers. Shaded nucleotides represent the nucleotides that were identical when the equivalent upstream regions of *C. trachomatis* D, *C. pneumoniae* AR39 and *C. psittaci* GPIC were aligned. The corresponding nucleotide/s in *C. trachomatis* D are marked, for TSSs determined in *C. muridarum* MoPn for CT098-rs1 and CT111-groES [23,31]; *C. trachomatis* serovar F for CT323-infA [25]; *C. psittaci* biovar 6BC for CT446-euo [35] and *C. trachomatis* L2 for CT062-tyrS, CT286-clpC, CT439-rs12, CT444-omcA [15], CT559-actJ, CT596-exbB, CT708-snf and CT752-efp2. The information content (IC) of the promoters was the sum of the IC scores for the individually predicted -35 and -10 hexamers. Mismatches are the maximum number of mismatches the predicted promoters in *C. pneumoniae* AR39 or *C. psittaci* GPIC were from the predicted promoter in *C. trachomatis* D. Putative product is as described by GenBank.

3.3. Footy performs better than the standard PWM method on *C. trachomatis*

To compare Footy to the standard PWM method [24], the upstream regions of *C. trachomatis* were analysed for promoters similar to *E. coli* σ^{70} promoters using the standard PWM method. The promoters and equations used to calculate the PWMs were the same as Footy and the model had a spacer length of 15–19 bp. The IC threshold of 10.0 bits (or 5.0 bits per hexamer) was determined by analysing the 536 upstream regions and the 122 downstream regions of *C. trachomatis* for promoters and false-positives, respectively, and by determining the statistical significance of the promoters predicted in the upstream regions. The statistical significance of the promoters predicted was determined using the χ^2 test on two contingency tables, one corresponding to the predictions in the upstream regions and the other the false-positives in the

downstream regions. At an IC threshold of 10.0 bits, 10 promoters were predicted in the upstream regions (Table S2, supplementary material). Of the 15 positive control promoters (described above) one was predicted correctly by the standard PWM method, the promoter of *infA*.

The 10 promoters predicted by the standard PWM method all showed high homology to the *E. coli* σ^{70} consensus sequence [17]. The number of mismatches that each predicted promoter deviated from the σ^{70} consensus sequence ranged from zero to two, with a statistical mode of two mismatches. The spacer between the -35 and -10 hexamers ranged from 16 to 18 bp, with a mode of 17 bp.

4. Discussion

Footy predicted 42 promoters that phylogenetically foot-printed across three species of *Chlamydia*. A computational

Table 3
Analysis of the 15 positive control promoters predicted and not predicted by Footy

Promoter name	True positives ^a	False negatives ^b		
		Promoter located partly or fully within a gene	Promoter eliminated by stage 1 of Footy ^c	Promoter eliminated by stage 2 of Footy ^c
<i>infA</i> (F)	✓			
<i>omcA</i> (L1,L2, 6BC, EAE and IOL207)		✓		
16S rDNA1 P1 (L2 and MoPn)			✓	✓
CT602 (L2)			✓	
<i>hctA</i> (L2 and MN)			✓	
<i>Itu A</i> (L2)			✓	✓
<i>Itu B</i> (L2)			✓	✓
<i>omp A</i> P1 (L2)			✓	✓
<i>omp A</i> P2-P3 (L2, GPIC and MN)			✓	✓
<i>pkn5</i> (L2)			✓	✓
<i>rpoD</i> (L2)			✓	
<i>sctU</i> (L2)			✓	
<i>secA</i> (L2)			✓	✓
<i>srp</i> (L1)			✓	✓
tRNAThr2 (F)				✓

Parentheses contain strain names. Strain abbreviation are as follows: 6BC, *C. psittaci* strain 6BC; EAE, *C. psittaci* strain EAE/A22/M; F, *C. trachomatis* serovar F; GPIC, *C. caviae* biovar GPIC; IOL207, *C. pneumoniae* strain IOL-207; L1, *C. trachomatis* serovar L1; L2, *C. trachomatis* serovar L2; MN, *C. psittaci* sub-species meningo-pneumonitis; and MoPn, *C. muridarum* biovar MoPn.

^a True positives refer to known promoters predicted by Footy.

^b False-negatives refer to known promoters not predicted by Footy.

^c Indicates at which stage (1 or 2) of Footy the positive control promoters were eliminated.

method that is capable of genome-wide phylogenetic footprinting of promoters across the multiple genome sequences of a bacterial genus has not been previously reported, thus demonstrating the usefulness of Footy.

Comparison of Footy to the standard PWM method shows that Footy performs better when analysing *Chlamydia* for σ^{66} promoters. The 42 promoters predicted by Footy contained up to five mismatches from the *E. coli* σ^{70} consensus sequence, whereas the standard PWM method predicted 10 promoters with up to two mismatches from the σ^{70} consensus sequence (Table 4). Footy used an IC threshold that was 7.0 bits lower than the IC threshold used by the standard PWM method in this study. This increases the likelihood of finding promoters.

The standard PWM method did, however predict three promoters (CT016, CT763 and *glyQ*) not predicted by Footy. This is because the standard PWM method keeps promoters that either do not have a homolog in the equivalent upstream regions of *C. pneumoniae* or *C. caviae*, or the equivalent upstream regions in *C. pneumoniae* or *C. caviae* cannot be identified.

Consequently, the standard PWM method analysed 231 more upstream regions of *C. trachomatis* than Footy. Therefore, the standard PWM method shows better sensitivity than Footy at IC thresholds above 10 bits when analysing *C. trachomatis* for σ^{66} promoters; hence Footy will not replace the standard PWM based programmes such as “ScanACE” (<http://arep.med.harvard.edu/mrnadata/mrnasoft.html>) [3] and “patser” [33]. However, Footy will be very effective when used in conjunction with these programmes to predict more promoters within an organism.

Eskin and colleagues [5] analysed 120 and 136 regions between divergently transcribed genes of *C. muridarum* and *C. pneumoniae*, respectively, for statistically significant overrepresented patterns homologous to the *E. coli* σ^{70} consensus sequence. The analysis used a promoter model of two hexamers separated by a 3–23 bp spacer using the programme MITRA. The authors reported that MITRA was unable to extract an overrepresented pattern from the upstream regions of *C. muridarum* that was homologous to the σ^{70} consensus sequence.

Table 4
Comparison of Footy to some of the preferred promoter prediction programmes in bacteria

Programme	Organism/s	Regions analysed	No. of predictions in genome/s	Positive controls predicted	Maximum no. of mismatches from the σ^{70} consensus sequence	No. also predicted by Footy
Standard PWM method	<i>C. trachomatis</i>	Upstream regions	10	1	2	7
MITRA	<i>C. muridarum</i>	Regions between divergently transcribed genes	0	N/A	N/A	N/A
MITRA	<i>C. pneumoniae</i>	Regions between divergently transcribed genes	27	N/A	3	1
Footy	<i>C. trachomatis</i> , <i>C. pneumoniae</i> , <i>C. caviae</i>	Equivalent upstream regions	42	1	5	N/A

When MITRA was applied to *C. pneumoniae* it discovered the overrepresented pattern, TTGACA N₁₉ ATAATT, which was made up of 27 hits. If the –10 hexamer of this overrepresented pattern is shifted 1 bp upstream, this pattern is identical at 11 of the 12 positions to the σ^{70} consensus sequence. Therefore, MITRA predicted 27 σ^{66} promoters in *C. pneumoniae*, some of which contained up to three mismatches from the σ^{70} consensus sequence. Interestingly, only one of these promoters (designated as RCPX0664_RCPX0066 [5]), was also predicted by Footy (CP0079) (Table 4).

The analysis of *Chlamydia* with the standard PWM method, MITRA and Footy show that by analysing equivalent upstream regions from multiple species for promoters that phylogenetically footprint, more promoters were predicted and some had fewer matches to the σ^{70} consensus sequence.

MITRA predicted promoters not predicted by Footy, because it is able to predict promoters without homologs in the other chlamydial species. Therefore, the performance of MITRA will not suffer if the promoters or associated genes are not conserved across the different bacterial genomes analysed. Pattern discovery programmes such as MITRA could also be used in conjunction with Footy to increase the number of promoters predicted within an organism.

Given the hypothesis that chlamydial promoters are expected to be well conserved in the equivalent upstream regions, the low number of promoters predicted as phylogenetically conserved across chlamydial species by Footy is surprising. Possible explanations are: (1) that homologous promoters may have been eliminated from the dataset of equivalent upstream regions because promoters in one or more species overlapped or were located within the coding regions of genes; (2) there was a low level of conserved σ^{66} promoters across different species of *Chlamydia*; and/or (3) that many of the σ^{66} promoters were dissimilar to the σ^{70} consensus sequence and therefore below the detection levels of Footy.

The major outcome of this study is the development of a new programme that predicts conserved promoters on a genome-wide scale across multiple bacteria, while performing equally as well or better than the current methods. For example, when analysing *Chlamydia*, Footy predicted more promoters with some having fewer matches to the *E. coli* σ^{70} consensus sequence and maintained a level of sensitivity and specificity comparable with other promoter prediction programmes. Finally, the increased number of σ^{66} promoters predicted by Footy in *Chlamydia* will be of significant value to researchers studying this organism.

Acknowledgments

We gratefully acknowledge Karl Eisler and Melinda Ziino from the Australian Genome Research Facility (Melbourne, Australia) for performing the fragment analysis on cDNA; Anthony Rasmussen from the High Performance Computing and Research Support (Queensland University of Technology, Brisbane, Australia) for his technical assistance and Michael Towsey from the Faculty of Information Technology

Innovation (Queensland University of Technology, Brisbane, Australia) for his technical assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.resmic.2007.08.005.

References

- [1] Azuma, Y., Hirakawa, H., Yamashita, A., Cai, Y., Rahman, M.A., Suzuki, H., Mitaku, S., Toh, H., Goto, S., Murakami, T., Sugi, K., Hayashi, H., Fukushi, H., Hattori, M., Kuhara, S., Shirai, M. (2006) Genome sequence of the cat pathogen, *Chlamydomonas felis*. DNA Res. 13, 15–23.
- [2] Belland, R.J., Zhong, G., Crane, D.D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W.L., Caldwell, H.D. (2003) Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. Proc. Natl. Acad. Sci. USA 100, 8478–8483.
- [3] Berg, O.G., von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. J. Mol. Biol. 193, 723–750.
- [4] Carlson, J.H., Porcella, S.F., McClarty, G., Caldwell, H.D. (2005) Comparative genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic strains. Infect. Immun. 73, 6407–6418.
- [5] Eskin, E., Keich, U., Gelfand, M.S., Pevzner, P.A. (2003) Genome-wide analysis of bacterial promoter regions. Pac. Symp. Biocomput. 29–40.
- [6] Gelfand, M.S. (1999) Recognition of regulatory sites by genomic comparison. Res. Microbiol. 150, 755–771.
- [7] Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucleic Acids Res. 28, 695–705.
- [8] Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., Mironov, A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. Brief Bioinform. 1, 357–371.
- [9] Geng, M.M., Schuhmacher, A., Muehldorfer, I., Bensch, K.W., Schaefer, K.P., Schneider, S., Pohl, T., Essig, A., Marre, R., Melchers, K. (2003). The Genome Sequence of *Chlamydia pneumoniae* TW183 and Comparison with Other *Chlamydia* Strains Based on Whole Genome Sequence Analysis. Byk Gulden Pharmaceuticals.
- [10] Gershenzon, N.I., Stormo, G.D., Ioshikhes, I.P. (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res. 33, 2290–2301.
- [11] Gordon, J.J., Towsey, M.W., Hogan, J.M., Mathews, S.A., Timms, P. (2006) Improved prediction of bacterial transcription start sites. Bioinformatics 22, 142–148.
- [12] Harley, C.B., Reynolds, R.P. (1987) Analysis of *E. coli* promoter sequences. Nucleic Acids Res. 15, 2343–2361.
- [13] Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., Stephens, R.S. (1999) Comparative genomes of *Chlamydia pneumoniae* and *Chlamydia trachomatis*. Nat. Genet. 21, 385–389.
- [14] Koehler, J.E., Burgess, R.R., Thompson, N.E., Stephens, R.S. (1990) *Chlamydia trachomatis* RNA polymerase major sigma subunit. Sequence and structural comparison of conserved and unique regions with *Escherichia coli* sigma-70 and *Bacillus subtilis* sigma-43. J. Biol. Chem. 265, 13206–13214.
- [15] Lambden, P.R., Everson, J.S., Ward, M.E., Clarke, I.N. (1990) Sulfur-rich proteins of *Chlamydia trachomatis*: developmentally regulated transcription of polycistronic mRNA from tandem promoters. Gene 87, 105–112.
- [16] Lee, P.S., Lee, K.H. (2000) Genomic analysis. Curr. Opin. Biotechnol. 11, 171–175.
- [17] Lissner, S., Margalit, H. (1993) Compilation of *E. coli* mRNA promoter sequences. Nucleic Acids Res. 21, 1507–1516.

- [18] Lonetto, M., Gribskov, M., Gross, C.A. (1992) The sigma-70 family: sequence conservation and evolutionary relationships. *J. Bacteriol.* 174, 3843–3849.
- [19] Mathews, S.A., Timms, P. (2000) Identification and mapping of sigma-54 promoters in *Chlamydia trachomatis*. *J. Bacteriol.* 182, 6239–6242.
- [20] Mathews, S.A., Timms, P. (2006) *In silico* identification of chlamydial promoters and their role in regulation and development. In P.M. Bavoil, & P.B. Wyrick (Eds.), *Chlamydia: Genomics and Pathogenesis*, Horizon Bioscience (pp. 133–156).
- [21] Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., Fraser, C.M. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28, 1397–1406.
- [22] Read, T.D., Myers, G.S., Brunham, R.C., Nelson, W.C., Paulsen, I.T., Heidelberg, J., Holtzapple, E., Khouri, H., Federova, N.B., Carty, H.A., Umayam, L.A., Haft, D.H., Peterson, J., Beanan, M.J., White, O., Salzberg, S.L., Hsia, R.C., McClarty, G., Rank, R.G., Bavoil, P.M., Fraser, C.M. (2003) Genome sequence of *Chlamydomydia caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the *Chlamydiaceae*. *Nucleic Acids Res.* 31, 2134–2147.
- [23] Sardinia, L.M., Engel, J.N., Ganem, D. (1989) Chlamydial gene encoding a 70-kilodalton antigen in *Escherichia coli*: analysis of expression signals and identification of the gene product. *J. Bacteriol.* 171, 335–341.
- [24] Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–431.
- [25] Shen, L., Shi, Y., Douglas, A.L., Hatch, T.P., O'Connell, C.M., Chen, J.M., Zhang, Y.X. (2000) Identification and characterization of promoters regulating *tuf* expression in *Chlamydia trachomatis* serovar F. *Arch. Biochem. Biophys.* 379, 46–56.
- [26] Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S., Nakazawa, T. (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res.* 28, 2311–2314.
- [27] Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282, 754–759.
- [28] Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23.
- [29] Stormo, G.D., Hartzell, G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* 86, 1183–1187.
- [30] Tan, M., Gaal, T., Gourse, R.L., Engel, J.N. (1998) Mutational analysis of the *Chlamydia trachomatis* rRNA P1 promoter defines four regions important for transcription *in vitro*. *J. Bacteriol.* 180, 2359–2366.
- [31] Tan, M., Wong, B., Engel, J.N. (1996) Transcriptional organization and regulation of the *dnaK* and *groE* operons of *Chlamydia trachomatis*. *J. Bacteriol.* 178, 6983–6990.
- [32] Thomson, N.R., Yeats, C., Bell, K., Holden, M.T., Bentley, S.D., Livingstone, M., Cerdeno-Tarraga, A.M., Harris, B., Doggett, J., Ormond, D., Mungall, K., Clarke, K., Feltwell, T., Hance, Z., Sanders, M., Quail, M.A., Price, C., Barrell, B.G., Parkhill, J., Longbottom, D. (2005) The *Chlamydomydia abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. *Genome Res.* 15, 629–640.
- [33] van Helden, J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.* 31, 3593–3596.
- [34] Wels, M., Francke, C., Kerkhoven, R., Kleerebezem, M., Siezen, R.J. (2006) Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res.* 34, 1947–1958.
- [35] Wichlan, D.G., Hatch, T.P. (1993) Identification of an early-stage gene of *Chlamydia psittaci* 6BC. *J. Bacteriol.* 175, 2936–2942.