Promoter prediction in the rhizobia

Shawn R. MacLellan, Allyson M. MacLean and Turlough M. Finan

Correspondence
Turlough M. Finan
finan@mcmaster.ca

Center for Environmental Genomics, Department of Biology, McMaster University, 1280 Main St West, Hamilton, Ontario L8S 4K1, Canada

The ability to recognize and predict non- σ^{54} promoters in the alphaproteobacteria is not well developed. In this study, 25 experimentally verified *Sinorhizobium meliloti* promoter sequences were compiled and used to predict the location of other related promoters in the *S. meliloti* genome. Fourteen candidate predictions were targeted for verification and of these at least 12 proved to be genuine promoters. As a result, the experimental identification of 12 novel promoters linked to genes *rpoD*, *topA*, *rpmJ*, *trpS*, *ropB1*, *metC*, *rpsT*, *secE*, *trkH* and three tRNA genes is reported. In all, 99 predicted and verified promoters are reported, including those linked with 13 tRNA genes, eight ribosomal protein genes and a number of other physiologically important or essential genes. On the basis of sequence conservation and a mutational analysis of promoter activity, the –35 and –10 consensus for these promoters is 5-CTTGAC-N₁₇-CTATAT. This promoter structure, which seems to be widely conserved amongst several other genera in the alphaproteobacteria, shares significant similarity with, but is skewed by a 1 nt step from, the canonical *Escherichia coli* σ^{70} promoter. Perhaps this difference is responsible for the observation that *S. meliloti* promoters are often poorly expressed in *E. coli*. In this regard, expression data from plasmid-borne *gfp*-reporter fusions to eight of the *S. meliloti* promoters verified in this work revealed that while these promoters were very active in *S. meliloti* and *Agrobacterium tumefaciens* only very low, near-background activity was detected in *E. coli*.

Received 4 December 2005

Revised 26 February 2006

Accepted 28 February 2006

INTRODUCTION

Sinorhizobium meliloti is one of a large group of alpha-proteobacteria that infect plant or animal cells and mediate symbiotic or pathogenic interactions within the host. In the case of *S. meliloti*, genes required for host invasion and nitrogen fixation are carried on a primary chromosome and two large (1.4 and 1.7 Mb) plasmids (Galibert *et al.*, 2001). For several members of this group, including *Agrobacterium tumefaciens*, *Brucella melitensis*, *B. suis*, *B. abortus*, *S. meliloti* and *Bradyrhizobium japonicum*, genome sequences have been published (DeVecchio *et al.*, 2002; Galibert *et al.*, 2001; Halling *et al.*, 2005; Kaneko *et al.*, 2002; Paulsen *et al.*, 2002; Wood *et al.*, 2001). As is increasingly evident in organisms such as *Escherichia coli*, the initial annotation of the genome sequence misses many genes and some of these (especially small RNA genes) have been found by scanning the genome for promoter-like elements linked to transcriptional termination elements (Argaman *et al.*, 2001; Chen *et al.*, 2002). Promoter prediction in *E. coli* is particularly well established, where the canonical consensus for σ^{70} promoters is based on sequence conservation amongst several hundred promoter sequences (Harley & Reynolds, 1987; Lissner & Margalit, 1993). Except for the σ^{54} promoters (Dombrecht *et al.*, 2002; Thony & Hennecke, 1989), the ability to predict and recognize promoter elements is poorly developed in most members of the alphaproteobacteria. We have therefore compiled a collection of experimentally verified *S.*

meliloti promoter sequences (Bae *et al.*, 1989; Fisher *et al.*, 1987; Gustafson *et al.*, 2002; Leong *et al.*, 1985; MacLellan *et al.*, 2005; Osteras *et al.*, 1995; Papp, 2004; Ronson *et al.*, 1987) and have used these to identify a large number of related sequences in the genome. We predict that these sequences represent novel RpoD-dependent promoter elements and have confirmed that 12 of them represent genuine active promoters. An analysis of experimentally verified *S. meliloti* promoter sequences indicates that (1) the –10 region is particularly diverse in sequence and poorly defined except for two highly conserved residues, and (2) the consensus –10 and –35 hexanucleotides share a significant degree of conservation with the *E. coli* RpoD-dependent promoter consensus but these sequences are positionally skewed relative to one another by a single nucleotide step.

METHODS

Bacterial strains, plasmids and oligonucleotides. The bacterial strains and plasmids used in this study are listed in Table 1. Both *S. meliloti* and *A. tumefaciens* were grown in LB broth supplemented with 100 µg streptomycin ml^{–1} and 30 µg gentamicin ml^{–1} (where appropriate) at 30 °C while *E. coli* was grown in the same medium supplemented with 5 µg gentamicin ml^{–1} at 37 °C. Antibiotic concentrations were doubled in solid medium. The oligonucleotides used for primer extension reactions are also listed in Table 1.

Sequence analysis. The first-generation promoter prediction was based upon an alignment of previously verified *S. meliloti* promoter

Table 1. Bacterial strains, plasmids and primer extension oligonucleotides used in this study

Strain or plasmid	Characteristic or genotype	Source or reference
Bacteria		
<i>E. coli</i> DH5 α	<i>supE44</i> Δ <i>lacU169</i> (ϕ 80 <i>lacZ</i> Δ M15) <i>hsdR17 recA1 endA1 gyrA96 thi-1 relA1</i>	Strain collection
<i>S. meliloti</i> Rm1021	Spontaneous Sm ^r isolate of <i>R. meliloti</i> SU47	Strain collection
<i>A. tumefaciens</i> At9023	pATC58 deletion derivative of strain C58 (Rif ^r Sm ^r)	Strain collection
Plasmids		
pOT1	<i>gfpUV</i> transcriptional reporter plasmid (Gm ^r)	Allaway <i>et al.</i> (2001)
pTH1560	162 bp PCR product encompassing <i>incA1</i> promoter cloned into pOT1	MacLellan <i>et al.</i> (2005)
pTH1985	Site-directed mutant (A $\rightarrow$ T in -35 promoter region) of pTH1560	This work
pTH1991	Site-directed mutant (C $\rightarrow$ T in -35 promoter region) of pTH1560	This work
pTH1982	189 bp PCR product encompassing <i>incA2</i> promoter cloned into pOT1	This work
pTH1983	Site-directed mutant (T $\rightarrow$ A in -35 promoter region) of pTH1982	This work
pTH1984	Site-directed double mutant (C $\rightarrow$ T, T $\rightarrow$ A in -35 promoter region) of pTH1982	This work
Oligonucleotides		
<i>topA</i>	GTATCTAGACGATTCCACCACTACAACATTC	
<i>secE</i>	GCGTCTAGATGCCATATAATGCTCACGC	
<i>rpoD</i>	GAGTCTAGAGTCCTGCCACTGAAACG	
<i>rpmE</i>	GATCTAGAGCAAGGCATTGCGACTTTGG	
<i>rpmJ</i>	GATCTAGATCCTGATCCGCAGGTTTCG	
<i>ptrB</i>	AGTCTAGACACGCAGCAGATTAAAGCG	
<i>metC</i>	TATCTAGACTTCCCGATCTTGTGTCTGC	
<i>trpS</i>	AGTCTAGATACGAGCGGCTTGAATTTCG	
<i>trkH</i>	AGTCTAGATAGTGCCGTGCGTGACC	
SMc02206 tRNA	AATCTAGAACCGACGACATCAAGCTTGG	
SMc01378 tRNA	AGTCTAGACTGTCGGGATCAATCGACC	
SMc00758 tRNA	AGTCTAGACTAACCTACTGAGCTACTCC	
<i>rpsT</i>	AGTCTAGACTGCGACCTGTCTACAAGC	
<i>ropB1</i>	AGTCTAGAGTACGCATACTATTCTCTTTCG	

sequences (Fig. 1, sequences 1–11). These sequences or sequence 10 (the *rRNA* operon promoter) were used to derive a search string that could be used to scan the *S. meliloti* genome for matches. A standard weight matrix was also derived based upon nucleotide frequency (percentage conservation) at each of the 30 nucleotide positions in the alignment of promoters. For the purposes of this initial investigation only linker regions (between the -10 and -35 hexanucleotides) of 17 nt were considered and in promoters possessing 18 nt linkers, a non-conserved position was eliminated in the linker sequence to facilitate construction of the matrix. The second-generation matrix was constructed as described for the first except that all 25 verified promoter sequences (Fig. 1) were used. The program PatScan (Dsouza *et al.*, 1997) was used to search the *S. meliloti* genome for matching sequences. In the case of the matrices, PatScan allows the input of a minimum threshold score such that only 30-mer sequences in the genome that match or exceed the score are reported. All sequences in this work were aligned manually.

Primer extension. Primer extension reactions were conducted using total bacterial RNA and oligonucleotides listed in Table 1 as previously described (MacLellan *et al.*, 2005). Extension and sequencing reactions were separated on 8% denaturing polyacrylamide gel containing 7 M urea. Signals were recorded on a phosphor screen and images were processed using ImageQuant v. 5.2.

Enzyme assays and other techniques. Estimates of promoter activity were obtained by cloning DNA fragments into the transcriptional reporter vector pOT1 (Allaway *et al.*, 2001) and measuring green fluorescent protein (GFP) fluorescence in a Tecan Safire fluorimeter. Cells were grown overnight then used to inoculate fresh

volumes of LB broth to an OD₆₀₀ of ~ 0.05 . Cultures were grown to an OD₆₀₀ of 0.7–0.8, washed once in 0.85% saline and resuspended in 1 vol. saline. Two hundred microlitre volumes were deposited in black microtitre plates and fluorescence was measured with an excitation wavelength of 405 nm and an emission wavelength of 505 nm. The optical densities of equivalent volumes were measured at 600 nm in clear microtitre plates. Specific fluorescence was obtained by dividing relative fluorescence by the optical density.

DNA sequencing and oligonucleotide synthesis was conducted by MOBIX lab (McMaster University, Hamilton).

RESULTS

Putative promoter sequences were predicted in two stages using a set of 11 experimentally verified promoter sequences from the existing literature in the first stage (first-generation prediction) and an expanded set of verified promoter sequences (consisting of an additional 14 promoter sequences verified in our laboratory) for the second stage (second generation) of predictions.

First-generation prediction of novel promoter sequences

Recently the characterization and mutational analysis of a promoter found upon the pSymB megaplasmid in *S. meliloti* led to the discovery of a small RNA gene that plays a role in

Promoter sequences			Gene		Reference
	-35	17-18 nt	-10		
1	CTTGAC	ACTGATT-CGCGGAAGTG	GGATTC	T	<i>incA1</i> (MacLellan et al. 2005)
2	CTTGAC	TGTGATT-CGTGGAAATG	CGATTC	T	<i>incA2</i> (MacLellan et al. 2005)
3	CTTGTC	TTGGGTC-AGCCTTGCCG	GTATGT	T	<i>pckA</i> (Osteras et al. 1995)
4	CTTGAC	CAAATTC-CAGTAATAAG	CAATTT	T	<i>ntrA</i> (Ronson et al. 1987)
5	CTTGAC	TTCCGATC-GATGTTCCGG	AGAATG	A	<i>hemA</i> P2 (Leong et al. 1985)
6	GTTGAC	CACTGAT-CGCTTTGAAG	GAAGAA	A	<i>hemA</i> P1 (Leong et al. 1985)
7	CTTGAT	TCCATTAACCTTCAGGGTT	CTCTAA	T	<i>nodD1</i> (Fisher et al. 1987)
8	CTTGAT	TTCGTAACTTCAGAGTT	CTCTAA	T	<i>nodD2</i> --
9	CTTGCG	CGCCAGC-GCAAGCCGCG	CTAACA	C	<i>trpE</i> (Bae et al. 1989)
10	GTTGAC	GTGTTGGAGGGCTGGGGT	CTATAA	G	<i>rRNA</i> (Gustafson et al. 2002)
11	CTTGTG	AAATCAA-AAATGCTGTT	CTATAG	A	tRNA-Pro (SMc00126) (Papp, 2004)
12	ATTGAT	GGCCATAGCCGAATCGGA	GTACGT	C	<i>repA2</i> --
13	ATTGAC	GCTTGCG-AAGAACAGGC	GCTTAT	T	<i>pcaI</i> --
14	GTTGAC	ACGATTG-GGTGACACG	GTATGT	G	<i>rpmJ</i> this work
15	CTTGAC	CGCGCCC-GAATCCCTGT	CCATGT	C	<i>topA</i> this work
16	CTTGAC	GGGATGA-AAAATTCTGG	GAATCA	C	<i>rpoD</i> this work
17	CTTGAC	CCGGCCG-GTGGCCCGTT	CGATGA	G	<i>metC</i> this work
18	CTTGAG	ATTCCTC-ATTTCTGAT	CAATTT	C	<i>ropB1</i> this work
19	CTTGAT	TGCGCCC-GGCCCGCCGT	GCATAA	G	<i>trpS</i> this work
20	GTTGAC	GTGCGGG-CGGTTTCCCT	TTATAC	G	<i>rpsT</i> this work
21	CTTGAC	ACTAGCCGGAACAGCGCC	TTATAA	G	tRNA-Met (SMc00758) this work
22	GTTGAC	AGGACGA-ATGCAGGCGG	TCATAT	A	tRNA-Tyr (SMc01378) this work
23	CTTGAC	AGATCACGGCTGTTGAGG	AAATTT	C	tRNA-Gly (SMc02206) this work
24	CTTGTG	GTTTTCG-CAATCGGTCT	TTATGT	A	<i>secE</i> this work
25	GTTGAG	TTTGCCG-ATGGCAGTCG	CTATTG	C	<i>trkH</i> this work
CTTGAC			CTATAT	c	consensus

Fig. 1. List of experimentally verified *S. meliloti* promoter sequences. Sequences 1–11 were identified from previously published sources and were used to derive search strings and a first-generation weight-based matrix that were employed in an initial attempt to predict novel promoters in the *S. meliloti* genome. Sequences 12 and 13 were obtained as a result of concurrent but independent investigations in our laboratory and sequences 14–25 were promoter sequences generated by our first-generation attempt to predict novel promoters that were subsequently verified using primer extension (see Fig. 4). The consensus sequence derived from the alignment is based upon the most abundant nucleotide at each position in the –10 and –35 hexanucleotide sub-sequences. Gaps (dashes) in the sequences were arbitrarily inserted to preserve alignment at the hexanucleotides. All 25 sequences were used to generate the second-generation weight-based matrix that is described in the text.

regulating the replication of the plasmid (MacLellan *et al.*, 2005). We noticed a striking conservation of sequence (CTTGAC) at the –35 region of the small RNA (*incA1*) promoter with other *S. meliloti* promoter sequences (Bae *et al.*, 1989; Fisher *et al.*, 1987; Leong *et al.*, 1985; Osteras *et al.*, 1995; Ronson *et al.*, 1987) previously reported in the literature. We therefore compiled 11 experimentally verified promoter sequences that were related to one another (most obviously at the –35 hexanucleotide sub-sequences) with the idea of using them to predict the location of other related promoter sequences in the *S. meliloti* genome. These 11 promoter sequences are listed (and aligned) in Fig. 1 (sequences 1–11). Two strategies were employed to identify sequences in the genome that had similarity to these sequences. First, we used the program PatScan (Dsouza *et al.*, 1997) to perform a string pattern search of the *S. meliloti* genome (see <http://sequence.toulouse.inra.fr/meliloti.html>)

using IUPAC symbols that constituted a degenerate consensus of the verified promoter sequences and a consensus based solely upon the rRNA operon promoter sequence. Amongst over 160 hits that were within ~250 bp on the proper strand upstream of a predicted ORF we identified two hits occurring upstream of genes involved in transcription (*suhR*, *rpoD*), twelve involved in translation [four tRNA genes (SMc01378, SMc01936, SMc00758, SMc00303), two aminoacyl synthetases (*cysS*, *trpS*) and six ribosomal proteins (*rluD*, *rpmE*, *rpmH*, *rpmJ*, *rplM*, *rpsT*)] and hits linked with a number of other physiologically important or essential genes including *secE* (protein translocation subunit), *topA* (topoisomerase I), two proteases (*ptrB*, SMc03769), *fixN3* (cytochrome *c* oxidase subunit), *expA1* (one of a 21-gene cluster required for exopolysaccharide synthesis), and a predicted chaperone gene, *grpE*. Eleven hits were linked with solute transporter loci and 14 were linked

with transcriptional regulatory protein genes. These hits are tabulated and aligned in Fig. 2.

We also constructed a weight-based matrix that we could use to assign scores to sub-sequences in the genome based on similarity to the verified promoter sequences. Best hits using this strategy were defined based on those sequences whose score equalled or exceeded an arbitrary discriminating threshold value. This first-generation weight matrix was used to scan the *S. meliloti* genome. The maximum score attainable with this matrix was 1860 and no sequence in the *S. meliloti* genome attained this score. Using a cutoff score of 1300, many more candidate promoter hits were obtained, including sequences linked to genes SMC02206 (tRNA),

cspA6 (a cold-shock protein), *dnaE1* (DNA polymerase subunit), *pip3* (proline iminopeptidase), *rpoZ* (RNA polymerase omega subunit) and *nuoA1* (the first gene in an 11-gene cluster encoding subunits of an NADH dehydrogenase complex) (Fig. 2).

The hits tabulated in Fig. 2 were selected from the first-generation prediction stage as their function is generally well established in bacteria and several are either very important or essential to cellular viability. These fall into well-defined physiological groups such as genes involved in translation or in ribosomal structure and function. The degree of sequence conservation within the -10 and -35 sub-sequences aligned in Fig. 2 is in most cases striking and

Sequences		Gene	Function
-35	-10		
1 CTTGTGTTGCAAAGAC-AATGCCGTAATCGC		<i>suhR</i>	transcription
2 CTTGACGGGATGAAAA-ATTCTGGGAATCAC		<i>rpoD</i>	
3 GATGGCAGCGTTCCAC-AGACGTGGCATT		<i>rpoZ</i>	translation
4 GTTGACAGGACGAATG-CAGGCGGTCATATA		tRNA (SMc01378)	
5 GTTGACCGGCGCAAAG-CCTGCGTGCATATG		tRNA (SMc00303)	
6 CTTGACACTAGCCGGAACAGCGCCTTATAAG		tRNA (SMc00758)	
7 GGTCGCTCCGCTGGGCCCGCATATATATAAG		tRNA (SMc01936)	
8 CTTGACAGATCACGGC-TGTTGAGGAAATTT		tRNA (SMc02206)	
9 CTTGACCGGCGCCGGG-CGACGCGCGATAAG		<i>cysS</i>	
10 CTTGATTGCGCCCGGC-CCGCCGTGCATAAG		<i>trpS</i>	
11 CTTGTCTTCTGTTCT-TCAGGCTCGATTGC		<i>rluD</i>	
12 CTTGTTATCCCGCAGA-CCTTCATGTATAGG		<i>rpmE</i>	
13 GTTGACACGATTGGGT-CGACACGGTATGTG		<i>rpmJ</i>	cytochrome oxidase exopolysaccharide protease protease protease outer membrane topoisomerase protein export drug resistance chaperone cold-shock protein NADH dehydrogenase PolIII subunit cystathionine beta-lyase potassium transporter
14 GTTGACGTGCGGGCGG-TTCCCTTTATACG		<i>rpsT</i>	
15 ATTGACTTTATCGGGTTTTTGTCCTATAAG		<i>rpmH</i>	
16 CTTGACCTTGTTTGGC-AGCTCCGTATAAG		<i>rpmM</i>	
17 CTTGATTTGTTCCCTC-GGGAGATAGATGAT		<i>fixN3</i>	
18 GTTGTGCACTGCGGAA-TCCACGGATATGGT		<i>expA1</i>	
19 CTTGATCGCCGCGCTG-GTGCCCTCTATCTC		<i>ptrB</i>	
20 CTTGTCTCCGGTCCGG-TCACAGGCCAACCT		<i>pip3</i>	
21 CTTGTCACGAAGCTCC-GAGTGTGCAATCTT		SMc03769	
22 CTTGAGATTCTTCATT-TCCTGATCAATTTT		<i>ropB1</i>	
23 CTTGACCGCGCCCGAA-TCCCTGTCCATGTC		<i>topA</i>	
24 CTTGTGGTTTTTCGCA-TCGGTCTTTATGTA		<i>secE</i>	
25 GTTGTTCGCCGGCCGT-TTTGATGCTATTAT		<i>aacC</i>	
26 CTTGATTTTTCGGGCC-CAACATCGATATC		<i>grpE</i>	
27 CTTGAGACCTCTTGCC-GTCCCGGGCATTGA		<i>cspA6</i>	
28 CTTGCGCCCGGGCGCC-GGCTGCGCTAACCA		<i>nuoA1</i>	
29 CTTGAGACTGGACAGT-TGGGACGGGAAACA		<i>dnaE1</i>	
30 CTTGACCGGCGGGTG-GCCCGTTTCGATGAG		<i>metC</i>	
31 GTTGAGTTTGCCGATG-GCAGTCGCTATTGC		<i>trkH</i>	

Fig. 2. Selected hits from the first-generation attempt to predict novel promoters. Sequences 3, 9, 22 and 29–31 were predicted on the basis of a high score obtained from a weight-based matrix derived from sequences 1–11 in Fig. 1. Sequences 7, 8 and 16–18 were predicted using a search string based upon the *rRNA* promoter sequence (see text). All of the remaining sequences were predicted using a search string analysis of the genome with a string based upon sequences 1–11 in Fig. 1. Hits shown were selected on the basis of their potential to be important or essential genes or because they cluster with other hits into functional groupings such as genes involved in translation or ribosomal structure. Some of these hits (sequences 2, 4, 6, 8, 10, 12, 13, 14, 19, 22, 23, 24, 30 and 31) were further examined to provide additional evidence that they represent bona fide promoter sequences.

transcriptional start sites for 14 of the genetic loci listed in Fig. 2 and examined whether these start sites correlated with the promoter elements we predicted for these genes. The targets we examined were *rpmJ*, *rpmE*, *ptrB*, *topA*, *rpoD*, *metC*, *ropB1*, *trpS*, *rpsT*, *secE*, *trkH*, and three tRNA genes (SMc00758, SMc01378, SMc02206). The reverse transcriptase products are shown in Fig. 4(a). No extension product was visible for *ptrB* (not shown). Several of the reactions (*rpmJ*, *topA*, *rpoD*, *ropB1*, SMc02206) generated multiple extension products but in every case (except for *rpoD*) the longest extension product (marked with an arrow) indicated a transcriptional start site that corresponds to the -10 and -35 regions that were predicted for each locus (Fig. 4b). In most cases, the shorter extension products (most of which are not shown due to space limitations) are probably due to premature termination of the reverse transcription reactions. In the case of *rpmE*, only a single extension product was observed but the transcriptional start site did not correspond to the predicted promoter element for this locus

(not shown). However, we note that the extension reaction in this case terminates near a GC-rich sequence that is annotated on the *S. meliloti* genome website as a *rho*-independent transcriptional terminator element. We therefore suspect that this product also represents a premature termination product. The distance between this GC-rich element and what may be the actual transcriptional start site (based on our promoter prediction) is only 17 nt and we therefore could not design a new primer to test this hypothesis. As shown in Fig. 3, however, the sequence we predict to be the *rpmE* promoter is positionally conserved in the related bacterium *A. tumefaciens*, suggesting that our prediction is most likely correct. In the case of *rpoD*, an extension product (see arrow) exactly correlates with the promoter we predicted for this locus but a longer extension product was also synthesized (not shown). The nucleotide sequence upstream of this additional start site was not obviously related to the confirmed promoter but nonetheless may represent an alternative promoter at this locus.

Fig. 4. Determination of transcriptional start sites upstream of target ORFs using primer extension. (a) Extension products using primers complementary to predicted transcripts from the gene as indicated. Arrows indicate the extension product that is consistent with the predicted promoter sequences for each locus. (b) Predicted promoter sequences (from Fig. 2) and the actual transcriptional start site (vertical arrow) for each locus determined from reactions in (a).

This experimental approach established that 12 of 13 ($\sim 92\%$) of the promoter elements we predicted and targeted for verification (not including *ptrB*, for which no information was obtained) are indeed genuine promoters and it is reasonable to suspect that the lone candidate that was not confirmed (*rpmE*) may be the result of premature termination of the primer extension reaction. In a circumstance that parallels the situation in *E. coli*, an A residue is the most frequently (50%) utilized transcription start nucleotide (Fig. 4b). The 12 predicted and confirmed promoter elements are listed in Fig. 1 (sequences 14–25) and all of the (now 25) verified promoter sequences were used to construct a second-generation weight matrix.

Second-generation weight-based matrix

We compiled the newly verified 12 promoter sequences (this work) with the original 11 sequences and with the two additional promoter sequences independently verified in our laboratory (sequences 12 and 13, Fig. 1) to generate a list of 25 verified promoter sequences (Fig. 1, sequences 1–25). Using these 25 verified promoter sequences we constructed a second-generation (and presumably more robust) weight matrix that we again used to scan intergenic regions in the *S. meliloti* genome using the program PatScan. The weight matrix in this case only identifies putative promoters that have a 17 nt linker region between the -35 and -10 sub-sequences.

The highest possible score with this matrix is 1484 and no sequence in the genome obtained this score. Five randomly generated 30-mer sequences (with the same G + C content as *S. meliloti* intergenic DNA: 56%) had a mean score of only 710 ± 59 (SE). Of the sequences used to construct the matrix (Fig. 1) the promoter for the *rRNA* gene has the highest score (1376) while the promoter for the *trpE* gene has the lowest score (1116). The highest scoring sequence (with a 17 nt linker) in the genome had a score of 1364 and was a hit linked to gene SMA0229, a hypothetical ORF whose predicted product has similarity to the translation elongation factor GreA. Hits linked to genes *cysS* and *topA* (both with a score of 1360) were the next highest scoring sequences in the genome. Using an arbitrary threshold score of >1190 , 411 hits were obtained. Ninety-five (23%) of these hits were not obviously linked to annotated genes. Some of these hits occurred in large intergenic regions while others occurred close to annotated genes but on the opposite strand. Seven of these orphan promoter-like sequences were closely linked (within ~ 200 nt) to orphan *rho*-independent transcriptional termination signals (that are annotated on the *S. meliloti* genome website), raising the possibility that together these elements may indicate the presence of small RNA genes. The number of orphan promoters closely linked with orphan terminators is likely an underestimation since a great many sequences that are good candidates for termination sequences are currently not annotated.

Eighteen and 12 hits, respectively, were closely linked with transcriptional regulatory protein genes and oxidoreductase/

dehydrogenase genes. Twenty-nine hits were linked with transporter genes. As befitting the frequency distribution of transporter types in the genome (Galibert *et al.*, 2001), most of the hits link to ABC transporter genes. These, along with other selected predictions, are listed in Table 2. Most of the hits listed in Table 2 come from the second-generation weight matrix analysis but some of the hits (including several tRNA gene and ribosomal protein gene promoter predictions) are duplications from Fig. 2 (from the first round of promoter predictions) and are included solely to facilitate sequence comparison amongst genes within a functional family. Table 2 also includes 32 hits linked to genes involved in translation and RNA metabolism and other hits to physiologically important or essential genes. Except for those few that we verified using primer extension, all of the sequences in Table 2 are of course putative but we expect roughly the same prediction success rate as demonstrated from the first generation of promoter predictions (that is 12 of 13 or minimally $\sim 90\%$).

Characterization of the *S. meliloti* promoter

Based upon the 25 verified *S. meliloti* promoter sequences (Fig. 1), we tabulated the frequency of each nucleotide at each of the positions delimited by the inferred -35 and -10 hexanucleotide sub-sequences and at each of 20 nucleotide positions upstream and downstream of the promoter (Fig. 5a). A plot of A + T frequency across the regions shows a defined profile of A/T-richness at the -10 and -35 sequences. The region upstream of the -35 hexanucleotide is also frequently rich in either A or T residues and 18 of the 25 sequences carry stretches of four or more contiguous A or T residues in the 20 nt that precede the -35 region. These sequences may function analogously to the UP elements found in the -40 to -60 region of many bacterial promoters (Aiyar *et al.*, 1998; Ross *et al.*, 1998). In contrast, the DNA region directly downstream of the -10 region is distinctly GC-rich, a situation reminiscent of the discriminator sequences found in some *E. coli* promoters (Zacharias *et al.*, 1990, 1991). A plot of conservation (percentage abundance) of the most abundant residue at each position (Fig. 5a) shows that the -35 region is well demarcated but that the -10 sub-sequence is ill-defined and more diverse than is the -35 sequence. The -35 region has the consensus sequence CTTGAC and this assignment is supported by the facts that the 5' and 3' terminal C residues are conserved in 68% and 64% of the sequences (respectively) (Fig. 5a) while the proximal flanking nucleotides on either end of the hexanucleotide show considerably less bias. We deduce that the -10 region has the consensus sequence CTATAT but except for nucleotides 3 and 4 (A,T), which are conserved in 88% and 80% of the sequences (respectively), there is less pronounced nucleotide bias at any of the positions (see *S. meliloti* sequence in Fig. 5b). The overall consensus (CTTGAC—CTATAT) specifies that 19 of the 25 verified promoters listed in Fig. 1 have a linker region that consists of 17 nt. The -35 and -10 *S. meliloti* sub-sequences bear striking similarity to the canonical *E. coli* promoter

Table 2. List of predicted *S. meliloti* promoters

Sequence	Score	Gene	Function
Transporters			
GTTGACGGTACTTGTA-CTAATATGTATAGA	1240	SMa0391	ABC transporter
CGTGACAGCTACGCCA-ACCCGTGGCATAGG	1200	SMa0451	ABC transporter
GTTGACGCATGTCAGG-AAACATTTTAGATC	1200	SMa0079	ABC transporter
GTTGACACTGTGAACT-TTTGATGCAACAAA	1196	SMa0953	ABC transporter
CTTGACGCTCGTTAGC-TTTGGTACTACCAG	1216	SMa1337	ABC transporter
CTTGACTGCTAATAAT-ATTAGCGCAAGGTT	1220	SMb21345	ABC transporter
GTTGACAGCTTGCGGA-AACTCTGACATTTA	1208	SMb20902	ABC transporter
CTTGACGGCGCCCGGA-TAAAATGGAATACA	1316	<i>mocB</i>	ABC transporter
GTTGACAAACGGACGG-TCTGAATCGATCAT	1204	SMb20155	ABC transporter
ATTGATTTGTGCTCTC-CGACGCGGAATAGT	1200	SMb20263	ABC transporter
GTTGACATTTGTCGAG-CCATGTTCGAATTA	1208	SMb20315	ABC transporter
ATTGACATGATTTGGG-AACCGGATCATCGT	1200	SMb20427	ABC transporter
CTTGACACATCTGGAC-TATTGAACAAGAAA	1204	SMb21221	ABC transporter
ATTGACATTTGAGCTC-AATCGGTGAGTAAT	1204	<i>smoE</i>	ABC transporter
CTTGGCAGCAGTGCCG-TCCTACTGCATGTT	1228	Y00717	ABC transporter
CTTGACACCGCGTCAG-TAGGATTTACTAAT	1252	Y03131	ABC transporter
CTTGACTCTTTCTCTG-CAGCGATGGAGTGT	1212	<i>aglE</i>	ABC transporter
ATTGCCCGGACCTCAG-ATCCGGGCAATTTT	1224	<i>potG</i>	ABC transporter
CTTGACCTTACCATTG-TTGAAGGACTATC	1232	<i>actP</i>	P-type Cu ²⁺ transporter
CTTGACCTTCCCATTA-TTGAAGGATTATG	1204	<i>actU2</i>	P-type Cu ²⁺ transporter
CTTGAATCACACGGGC-TTTCGCGGTATAAC	1236	SMa1664	Multi-drug efflux resistance
TTTGACTTTGGTCATC-CGCTGGTCAATTTC	1200	SMa1691	<i>trkH</i> -like cation transporter
CTTGATTCAAAGCAAG-TTGAGTGGAAGGAT	1196	<i>kefB2</i>	K ⁺ efflux
CTTGACCCGGGATTCA-GCGTCCCTATTTTC	1200	SMb20361	Voltage-gated transporter
CTTGAGGCAGGCCCGC-CCTTCTTCTATGTC	1272	<i>phaA2</i>	Cation efflux
CTTGGCAGCCGTCCGG-AAAATAGTCATAAA	1200	<i>phaA1</i>	Cation efflux
CTTGACGGGAATCGAA-TACTGGAGAACTCT	1212	<i>aqpZ1</i>	Aquaporin
ATTGATTTTGCGCCGG-GGTTGCGTCATATC	1236	<i>corA2</i>	Mg ²⁺ , Co ²⁺ transporter
CTTGACGGAACCGCGC-GCCGGCGCTCGAAT	1220	SMc04404	Amino acid efflux
RNA metabolism			
CTGGAAGAGAATCATG-CCCGGTGGCATGTC	1192	SMa0022	Sec-tRNA (TGA)
CTTGACACTAGCCGGAACAGCGCCTTATAAG	1344	SMc00758	Met-tRNA (CAT)
GGTCGCTCCGCTGGGCCCGCATATATATAAG	984	SMc01936	Asp-tRNA (GTC)
GTTGACAGGACGAATG-CAGGCGGTCATATA	1284	SMc01378	Tyr-tRNA (GTA)
CTTGAAAGCGTCCGGC-AAACTGAATATTAA	1224	SMc01381	Trp-tRNA (CCA)
CTTGACTGAATCGCCT-ATGTGCCTGATTTA	1224	SMc00957	Leu-tRNA (TAA)
GTTGACCGGCGCAAAG-CCTGCGTGCATATG	1276	SMc00303	Lys-tRNA (TTT)
CTTGCGGGAATCCGAA-GCCCTCTCTATAAG	1224	SMc00304	Glu-tRNA (TTC)
CTTGACATGAACCCGAT-GCCCTCTCTATAAG	1212	SMc00305	Glu-tRNA (CTC)
CTTGCGGAAATCCGAA-GCGCTCTCTATAAG	1228	SMc00306	Glu-tRNA (TTC)
CTTGCCAGCGGCAGAA-ACTGCCGCTATGAA	1264	SMc03779	Ser-tRNA (CGA)
CTTGAAATCTCGCCGC-TTTGCGTTTATCAG	1264	SMc02183	Thr-tRNA (CGT)
CTTGACAGATCACGGC-TGTTGAGGAAATTT	1208	SMc02206	Gly-tRNA (CCC)
CTTGACCGGCGCCGGG-CGACGCGCGATAAG	1360	<i>cysS</i>	tRNA synthetase
CTTGACAAAACCGGCG-CATCATGCGCTTTT	1212	<i>aspS</i>	tRNA synthetase
GTTGACTTGCCCCGGG-CCTTCCGGTTTACA	1200	<i>valS</i>	tRNA synthetase
CTTGACCTTCGCCCTG-CGGCATGGCACCTG	1208	<i>glyQ</i>	tRNA synthetase
CTTGATTGCGCCCGGC-CCGCCGTGCATAAG	1316	<i>trpS</i> (P1)	tRNA synthetase
CTTGTTGAGGGCCCGG-CTCCAGTTGATATC	1220	<i>trpS</i> (P2)	tRNA synthetase
CTTGCCGTGACGGGCG-CAAGTTTCTATATG	1272	<i>trmU</i>	tRNA methyltransferase
CTTGCTCTTCCTGTTCT-TCAGGCTCGATTGC	1168	<i>rluD</i>	Ribosomal protein
CTTGTTATCCCGCAGA-CCTTCATGTATAGG	1172	<i>rpmE</i>	Ribosomal protein

Table 2. cont.

Sequence	Score	Gene	Function
ATTGACTTTATCGGGTTTTTGTCCCTATAAG	1284	<i>rpmH</i>	Ribosomal protein
CTTGACCTTGTGCG-AGCTTCCGTATAAG	1284	<i>rplM</i>	Ribosomal protein
GTTGACACGATTGGGT-CGACACGGTATGTG	1312	<i>rpmJ</i>	Ribosomal protein
CTTGCCTTTTCCGGGA-TGTGCGGTATAAG	1276	<i>rpmI</i>	Ribosomal protein
GTTGACGTGCGGGCGG-TTTCCTTTATACG	1264	<i>rpsT</i>	Ribosomal protein
CTTGACTGCGGGGCA-CCTGTGACTATAGA	1272	<i>rpmB</i>	Ribosomal protein
CTTTACATCGGCGGCG-GGCATCGCTATTTG	1228	<i>prfB</i>	Peptide chain release factor
CTTGCGGCGTTTTACG-TGAATCGCTATTTT	1212	<i>rho</i>	Transcript termination protein
CTTGACTTTCGGCGCC-TTTGGTCGTATGAG	1308	<i>rhIE1</i>	RNA helicase
CTTGACTTTCGGCGCG-AAAACAACGATCTA	1200	<i>rhIE2</i>	RNA helicase
Others			
GTTGATGCCGCTGCGG-TCAAGCGCACTGAC	1216	SMa0116	Probable DnaJ (chaperone)
CTTGAATCAGAGCTTC-CATAGGTCTATACA	1236	<i>dskA</i>	DnaK suppressor
CTTGAACTCGTGCGCG-GGCATTCCCATGTC	1200	<i>ibpA</i>	Heat-shock (hsp70)
CTTGATTTTTTCGGGCC-CAAACATCGATATC	1224	<i>grpE</i>	Heat-shock
CTTGACTCATTTACG-GCCAGCCTTATCTC	1276	<i>groES1</i>	Chaperone
CTTGGCGTCCGCGAGG-ACCTCAACTATAAG	1200	<i>exsA</i>	Exopolysaccharide synthesis
AGTGACATCTGCGATG-TCAATGTCCATGAG	1156	<i>exoP2</i>	Exopolysaccharide synthesis
CTTGACGCCGTGCCGA-TAACCGTTCCTAAT	1252	<i>thiC</i>	Thiamine biosynthesis
GTTGCCTCCCTGCGAG-GCGGTGCGCTATGTT	1204	<i>sodB</i>	Superoxide dismutase
CTTGATTTTGTGCTGCTC-TTGCTTCTCTGTG	1240	<i>ftsZ2</i>	Z-ring cell division protein
CTTGACTCAATCGTAT-GAACTGACAATCAA	1216	<i>gidA</i>	Cell division protein
CTTGATTTTCGCGCCTG-CCCGACGCCACTGA	1200	<i>mrCA1 (P1)</i>	Penicillin-binding protein 1A
GTTGATGCTTCAGGTT-CCTGGCCATATGAG	1200	<i>mrCA1 (P2)</i>	Penicillin-binding protein 1A
CTTGTGGTTTTTCGCAA-TCGGTCTTTATGTA	1212	<i>secE</i>	Protein transport
GTTGACATAATCCGAC-CCCCCTCTTATCTG	1216	<i>secA</i>	Protein transport
CTTCACGCCGGGCGGT-TTGTCTCCCATATG	1204	<i>secY</i>	Protein transport
CTTGACTTTTCTTCCG-GTTTCCTGTTTATC	1248	<i>ffh</i>	Signal recognition particle
CCTGACAGCATCGCTA-CATCGCGTAATCAC	1236	SMc00168	LuxI autoinducer synthase
ATGGACCTTTCCCTG-CCTCGCGCTATCCA	1200	<i>polA</i>	DNA polymerase I
CTTCACTTTCGACCTG-ACTCTGTCCATATT	1224	<i>uvrB</i>	Nucleotide excision repair
ATTGAAAGCCGCGCG-CCAGAAGTTATATC	1192	<i>irr</i>	Fur homologue
GTTGACTGGCGGCCT-GCTGGCGGCAGAAA	1192	SMc00784	Iron-binding protein
CGTGACGGCTCCGAGA-CACCGAAGGATCAC	1196	<i>cobG</i>	Vitamin B12 biosynthetic cluster
CTTGATTTTCGAGCTCG-GCTCGGTGCATAAG	1280	<i>thdF</i>	Thiophene and furan oxidation
ATTGACTTCCGCGGT-ACAGCGGATATGAG	1240	SMc01140	σ^{54} modulation protein
GTTGCCGCTTTCCGG-AATGGTCCGATTTC	1220	<i>flaC</i>	Flagellin
CTTGACTATGCGGGTG-CCTAAACTATGTT	1296	<i>atp1</i>	F ₀ F ₁ ATP synthase cluster

consensus (Harley & Reynolds, 1987; Lissner & Margalit, 1993) and these sequences are directly compared in Fig. 5(b). At the -35 sub-sequence, both *E. coli* and *S. meliloti* promoters share a core sequence of TTGAC but they are shifted or skewed relative to one another by a 1 nt step, at least as evidenced by the conservation of nucleotides in these regions (Fig. 5b).

Given the significant similarity between these *S. meliloti* promoters and the canonical *E. coli* promoter, we cloned approximately 200 bp regions from upstream of the genes *rpoD*, *secE*, *rpmJ*, *rpmE*, *ropB1*, SMc1378, *rpsT* and *topA* (see list in Fig. 2) into the *gfp* transcriptional reporter pOT1 and promoter activity was measured in *E. coli*, *A. tumefaciens*

and *S. meliloti*. As a control for these experiments, we used the *incA* promoter, which is known to be expressed in *E. coli* (MacLellan *et al.*, 2005). Most of the promoters were highly active in both *A. tumefaciens* and *S. meliloti*, but other than the control *incA*, none were appreciably active in *E. coli*. In this host, the most active promoter belonged to *rpmE* (with activity 1.9-fold above background) (Table 3). We expect that several of these housekeeping genes encode factor-independent promoters and thus should technically be capable of activity in heterologous hosts. In some instances, however, promoter activity may be dependent on host-specific transcriptional regulators not present in *E. coli*. Nevertheless, this survey is quite consistent with the common finding in our laboratory that *S. meliloti* promoters

Fig. 5. Analysis of the *S. meliloti* promoter. (a) Frequency of most abundant nucleotide at each position in 25 experimentally verified promoters (see Fig. 1). The sequence plotted ranges from 20 nt upstream of the –35 sub-sequence to 20 nt downstream of the –10 sub-sequence. The most abundant nucleotide(s) at each position is labelled (–35 and –10 nucleotides are capitalized). A+T abundance at each position is shown. Reference lines indicate the mean genomic and intergenic region A+T content. (b) Nucleotide frequency at each *E. coli* and *S. meliloti* –35 and –10 position as deduced from sequences 1–25 in Fig. 1. Consensus nucleotides are in bold capitals; flanking nucleotides are in lower case.

rarely display activity in *E. coli*, a circumstance that provided a major motivation towards better defining at least a subset of *S. meliloti* promoter sequences.

Despite the conservation of core sequence (TTGAC) between the *E. coli* and *S. meliloti* –35 hexanucleotides we have concluded that the *S. meliloti* consensus is completed by a 5' C residue (CTTGAC) as opposed to the *E. coli* consensus, which is completed by a 3' A residue (TTGACA) based on the frequency of conserved nucleotides. Using site-directed mutagenesis on two *S. meliloti* promoters (the *incA1* and *incA2* gene promoters) we sought to provide experimental support for the idea that the 5' C residue of the consensus is functionally important for promoter activity and, conversely, that the 3' nucleotide most proximal to CTTGAC has little functional importance. To do this, we cloned both promoter regions into the *gfp* transcriptional reporter vector pOT1 and monitored promoter activity as specific fluorescence in both *A. tumefaciens* and *E. coli* cells. We found it necessary to use the closely related

bacterium *A. tumefaciens* as host for the test plasmids because the products expressed from the *incA1* and *incA2* promoters (~56 nt untranslated RNAs) mediate incompatibility against the resident megaplasmids of *S. meliloti* (MacLellan *et al.*, 2005). We know from this and previous work that patterns of expression in *A. tumefaciens* and *S. meliloti* are comparable.

As shown in Table 4, the *incA1* (pTH1560) and *incA2* (pTH1982) promoters are highly active in both *A. tumefaciens* and *E. coli*. Mutations in the nucleotide 3' of the consensus (pTH1985 and pTH1983, respectively) have little or no impact on promoter activity in both species. We expected this result for *A. tumefaciens* cells since that nucleotide falls outside the consensus hexanucleotide and is less biased than the other adjacent nucleotides. In the case of *E. coli*, however, we found this surprising since we expected that a T in that nucleotide position (which is found much less frequently than an A in that position) might lower activity if *E. coli* RNA polymerase continued to recognize

Table 3. Activity of selected *S. meliloti* promoters in parental or heterologous host strains

Promoter*	Mean promoter activity $\pm$ SD (fold above background)†		
	<i>E. coli</i>	<i>A. tumefaciens</i>	<i>S. meliloti</i>
pOT1	307 $\pm$ 14 (1)	168 $\pm$ 5 (1)	115 $\pm$ 7 (1)
<i>rpoD</i>	478 $\pm$ 4 (1.6)	1481 $\pm$ 49 (8.8)	4405 $\pm$ 9 (38.3)
<i>secE</i>	520 $\pm$ 20 (1.7)	8971 $\pm$ 95 (53.4)	25003 $\pm$ 253 (217.4)
<i>rpmJ</i>	237 $\pm$ 16 ($\leq$ 0)	14973 $\pm$ 1302 (89.1)	32452 $\pm$ 1091 (282.2)
<i>rpmE</i>	592 $\pm$ 4 (1.9)	4101 $\pm$ 131 (24.4)	14833 $\pm$ 276 (129.0)
<i>ropB1</i>	373 $\pm$ 9 (1.2)	984 $\pm$ 66 (5.9)	1803 $\pm$ 44 (15.7)
<i>SMc1378</i>	405 $\pm$ 12 (1.3)	19358 $\pm$ 685 (115.2)	82752 $\pm$ 2929 (719.6)
<i>rpsT</i>	426 $\pm$ 17 (1.4)	9665 $\pm$ 84 (57.5)	48076 $\pm$ 1643 (418.1)
<i>topA</i>	433 $\pm$ 24 (1.4)	1567 $\pm$ 55 (9.3)	2854 $\pm$ 63 (24.8)
<i>incA</i> ‡	34335 $\pm$ 405 (111.8)	1986 $\pm$ 73 (11.8)	NA§

*The test plasmid was pOT1 or pOT1 carrying the indicated gene promoter (see text).

†GFP activity values reflect data from three replicates in the indicated hosts. Background activity is from cells carrying the reporter vector, pOT1.

‡The DNA fragment containing the small RNA gene *incA* promoter also carries the transcriptional terminator for the gene. This terminator is poorly recognized in *E. coli* but efficiently terminates in *A. tumefaciens*. Thus, the activity for the *incA* promoter in *A. tumefaciens* will be about 10–20 % of the actual promoter activity.

§NA, Not applicable. Since expression of the full-length *incA* gene mediates incompatibility in *S. meliloti*, the *incA* promoter was not tested in this background.

the *E. coli* consensus hexanucleotide that is incidentally formed in plasmids pTH1560 and pTH1983 but not plasmids pTH1985 and pTH1982. In all cases, when the inferred 5' nucleotide of the *S. meliloti* consensus (CTTGAC) was mutated to form TTGAC (plasmids pTH1991 and pTH1984), promoter activity was significantly affected. For reasons we have not determined, the influence of the mutation is much more dramatic in the *incA2* promoter

than in the *incA1* promoter. Although the *incA1* and *incA2* promoter regions are highly similar, sequence differences elsewhere in the promoters are presumably responsible for this effect. In any case, the weight matrix analysis suggests that the –35 hexanucleotide rarely if ever begins with a T residue and this is consistent with the dramatic loss of promoter activity from promoters that carry this nucleotide in that position.

Table 4. Mutational analysis of *S. meliloti* promoter –35 sub-sequence

Promoter*	Plasmid	Mutation	Sequence	–35 consensus n T T G A C A C T T G A C n	
				Activity† (× 1000)	
				<i>A. tumefaciens</i>	<i>E. coli</i>
<i>incA1</i>	pOT1	–	–	0.79 (0.09)	1.03 (0.07)
	pTH1560 (wt)	None	C T T G A C a	52.4 (2.7)	72.2 (1.7)
	pTH1985	a→t	C T T G A C t	52.7 (1.9)	70.0 (1.7)
<i>incA2</i>	pTH1991	C→T	T T T G A C a	15.8 (0.3)	35.5 (1.6)
	pTH1982 (wt)	None	C T T G A C t	49.7 (0.2)	69.7 (0.9)
	pTH1983	t→a	C T T G A C a	38.9 (0.1)	73.4 (1.0)
	pTH1984	T→a, C→T	T T T G A C a	1.5 (0.05)	8.1 (0.06)

*Gene promoter region (as depicted in Fig. 1) cloned into transcriptional reporter vector pOT1 (as described in Table 1).

†Activity measured as specific GFP fluorescence and reported as the mean of triplicate readings ($\pm$ SE). *A. tumefaciens* was used as a surrogate host for the plasmids as some retain the ability to mediate incompatibility against the resident megaplasmids in *S. meliloti* (see Methods).

DISCUSSION

The ability to predict promoter elements in the alpha-proteobacteria is only now emerging as the result of the accumulation of experimentally verified promoter sequences and access to the genomic sequences of these bacteria. Except for the σ^{54} promoters (Dombrecht *et al.*, 2002), we are aware of no concerted effort to predict promoters in any members of the group. Our effort represents a fundamental first step in this direction and as a result of this work we report the experimentally verified sequences of an additional 12 *S. meliloti* promoters. As more sequences are verified, this work can be extended and refined. We provide a list (Table 2) of nearly 100 predicted promoters but these represent only a fraction of sequences in the *S. meliloti* genome that appear to be related to the consensus we have defined in this work.

The structure demonstrated for *S. meliloti* promoters appears to be conserved amongst other bacteria in the alphaproteobacteria for the following reasons: (1) we showed that the promoter sequences for the *rpmE*, *rpoD*, *topA* and *secE* genes were conserved between *S. meliloti* and *A. tumefaciens* (Fig. 3); (2) homologues of the *incA1* and *incA2* genes whose promoters have been verified have homologues in *A. tumefaciens*, *Rhizobium etli*, *R. leguminosarum* and *Brucella* spp. (Chai & Winans, 2005; Dombrecht *et al.*, 2002; Izquierdo *et al.*, 2005; MacLellan *et al.*, 2005; Venkova-Canova *et al.*, 2004) and their promoters are nearly identical; and (3) a small compilation (eight genes) of *Bradyrhizobium japonicum* promoters (Beck *et al.*, 1997) demonstrated the same general pattern of nucleotide conservation as described in this work.

We have not experimentally deduced which sigma factor recognizes the confirmed and putative promoters that have been predicted in this work. However, the sigma factor is likely to be *rpoD*, the vegetative sigma factor, since many of the confirmed and predicted loci associated with these promoters encode products required for translation and transcription, and other essential or important functions.

In this work we have identified consensus -10 and -35 hexameric promoter sequences defining what we believe to be a subset of *S. meliloti* promoters. Further experimental work will be required to better define these sequences from a functional point of view and to determine whether the *E. coli* paradigm of conserved hexamers is appropriate for *S. meliloti* promoters, particularly with regard to the poorly conserved -10 region. From the perspective of predicting related sequences (putative promoters) in the alphaproteobacteria, however, the identification of patterns of nucleotide conservation described in this work is a productive first step towards more effective promoter prediction in these bacteria.

ACKNOWLEDGEMENTS

This work was supported with funding from Genome Canada through the Ontario Genomics Institute and with funding from the Ontario

research and development challenge fund. The work was initially supported by a grant from the Natural Sciences and Engineering Research Council to T.M.F. We are grateful to Marie Elliot, Alison Cowie and Richard Morton for input and comments on the manuscript.

REFERENCES

- Aiyar, S. E., Gourse, R. L. & Ross, W. (1998). Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase alpha subunit. *Proc Natl Acad Sci U S A* **95**, 14652–14657.
- Allaway, D., Schofield, N. A., Leonard, M. E., Gilardoni, L., Finan, T. M. & Poole, P. S. (2001). Use of differential fluorescence induction and optical trapping to isolate environmentally induced genes. *Environ Microbiol* **3**, 397–406.
- Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. & Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol* **11**, 941–950.
- Bae, Y. M., Holmgren, E. & Crawford, I. P. (1989). *Rhizobium meliloti* anthranilate synthase gene: cloning, sequence, and expression in *Escherichia coli*. *J Bacteriol* **171**, 3471–3478.
- Beck, C., Marty, R., Klausli, S., Hennecke, H. & Gottfert, M. (1997). Dissection of the transcription machinery for housekeeping genes of *Bradyrhizobium japonicum*. *J Bacteriol* **179**, 364–369.
- Chai, Y. & Winans, S. C. (2005). A small antisense RNA down-regulates expression of an essential replicase protein of an *Agrobacterium tumefaciens* Ti plasmid. *Mol Microbiol* **56**, 1574–1585.
- Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J. & Blyn, L. B. (2002). A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems* **65**, 157–177.
- DeVecchio, V. G., Kapatral, V., Elzer, P., Patra, G. & Mijer, C. V. (2002). The genome of *Brucella melitensis*. *Vet Microbiol* **90**, 587–592.
- Dombrecht, B., Marchal, K., Vanderleyden, J. & Michiels, J. (2002). Prediction and overview of the RpoN-regulon in closely related species of the *Rhizobiales*. *Genome Biol* **3**, RESEARCH0076.
- Dsouza, M., Larsen, N. & Overbeek, R. (1997). Searching for patterns in genomic data. *Trends Genet* **13**, 497–498.
- Fisher, R. F., Brierley, H. L., Mulligan, J. T. & Long, S. R. (1987). Transcription of *Rhizobium meliloti* nodulation genes. Identification of a *nodD* transcription initiation site *in vitro* and *in vivo*. *J Biol Chem* **262**, 6849–6855.
- Galibert, F., Finan, T. M., Long, S. R. & 53 other authors (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668–672.
- Gustafson, A. M., O'Connell, K. P. & Thomashow, M. F. (2002). Regulation of *Sinorhizobium meliloti* 1021 *rrnA*-reporter gene fusions in response to cold shock. *Can J Microbiol* **48**, 821–830.
- Halling, S. M., Peterson-Burch, B. D., Bricker, B. J., Zuerner, R. L., Qing, Z., Li, L. L., Kapur, V., Alt, D. P. & Olsen, S. C. (2005). Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *J Bacteriol* **187**, 2715–2726.
- Harley, C. B. & Reynolds, R. P. (1987). Analysis of *E. coli* promoter sequences. *Nucleic Acids Res* **15**, 2343–2361.
- Izquierdo, J., Venkova-Canova, T., Ramirez-Romero, M. A., Tellez-Sosa, J., Hernandez-Lucas, I., Sanjuan, J. & Cevallos, M. A. (2005). An antisense RNA plays a central role in the replication control of a *repC* plasmid. *Plasmid* **54**, 259–277.

- Kaneko, T., Nakamura, Y., Sato, S. & 14 other authors (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110 (supplement). *DNA Res* 9, 225–256.
- Leong, S. A., Williams, P. H. & Ditta, G. S. (1985). Analysis of the 5' regulatory region of the gene for delta-aminolevulinic acid synthetase of *Rhizobium meliloti*. *Nucleic Acids Res* 13, 5965–5976.
- Lisser, S. & Margalit, H. (1993). Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res* 21, 1507–1516.
- MacLellan, S. R., Smallbone, L. A., Sibley, C. D. & Finan, T. M. (2005). The expression of a novel antisense gene mediates incompatibility within the large repABC family of alpha-proteobacterial plasmids. *Mol Microbiol* 55, 611–623.
- Osteras, M., Driscoll, B. T. & Finan, T. M. (1995). Molecular and expression analysis of the *Rhizobium meliloti* phosphoenolpyruvate carboxykinase (*pckA*) gene. *J Bacteriol* 177, 1452–1460.
- Papp, P. P. (2004). tRNA gene containing an attachment site of a temperate phage is functional: proof by converting *attB* to amber suppressor. GenBank accession no. AJ698943.
- Paulsen, I. T., Seshadri, R., Nelson, K. E. & 28 other authors (2002). The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc Natl Acad Sci U S A* 99, 13148–13153.
- Ronson, C. W., Nixon, B. T., Albright, L. M. & Ausubel, F. M. (1987). *Rhizobium meliloti ntrA* (*rpoN*) gene is required for diverse metabolic functions. *J Bacteriol* 169, 2424–2431.
- Ross, W., Aiyar, S. E., Salomon, J. & Gourse, R. L. (1998). *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. *J Bacteriol* 180, 5375–5383.
- Thony, B. & Hennecke, H. (1989). The $-24/-12$ promoter comes of age. *FEMS Microbiol Rev* 5, 341–357.
- Venkova-Canova, T., Soberon, N. E., Ramirez-Romero, M. A. & Cevallos, M. A. (2004). Two discrete elements are required for the replication of a repABC plasmid: an antisense RNA and a stem-loop structure. *Mol Microbiol* 54, 1431–1444.
- Wood, D. W., Setubal, J. C., Kaul, R. & 49 other authors (2001). The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294, 2317–2323.
- Zacharias, M., Goring, H. U. & Wagner, R. (1990). The signal for growth rate control and stringent sensitivity in *E. coli* is not restricted to a particular sequence motif within the promoter region. *Nucleic Acids Res* 18, 6271–6275.
- Zacharias, M., Theissen, G., Bradaczek, C. & Wagner, R. (1991). Analysis of sequence elements important for the synthesis and control of ribosomal RNA in *E. coli*. *Biochimie* 73, 699–712.