

# A phylogenetic analysis of the pSymB replicon from the *Sinorhizobium meliloti* genome reveals a complex evolutionary history

Kim Wong and G. Brian Golding

**Abstract:** Microbial genomes are thought to be mosaic, making it difficult to decipher how these genomes have evolved. Whole-genome nearest-neighbor analysis was applied to the *Sinorhizobium meliloti* pSymB replicon to determine its origin, the degree of horizontal transfer, and the conservation of gene order. Prediction of the nearest neighbor based on contextual information, i.e., the nearest phylogenetic neighbor of adjacent genes, provided useful information for genes for which phylogenetic relationships could not be established. A large portion of pSymB genes are most closely related to genes in the *Agrobacterium tumefaciens* linear chromosome, including the *rep* and *min* genes. This suggests a common origin for these replicons. Genes with the nearest neighbor from the same species tend to be grouped in "patches". Gene order within these patches is conserved, but the content of the patches is not limited to operons. These data show that 13% of pSymB genes have nearest neighbors in species that are not members of the *Rhizobiaceae* family (including two archaea), and that these likely represent genes that have been involved in horizontal transfer.

**Key words:** *Sinorhizobium meliloti*, horizontal transfer, pSymB evolution.

**Résumé:** Les génomes microbiens sont reconnus comme étant mosaïques, ce qui rend difficile la tâche de retracer l'évolution de leur génome. Une analyse du plus proche voisin englobant le génome entier du réplikon pSymB de *Sinorhizobium meliloti* a été effectuée afin de déterminer son origine, le degré de transfert horizontal et la conservation de l'ordre des gènes. Nous avons démontré l'utilité des prédictions du plus proche voisin basées sur l'information contextuelle, c.-à-d. le voisin le plus proche de gènes adjacents, afin de fournir des informations précieuses à propos de gènes dont les relations phylogénétiques n'avaient pu être établies. Une part importante des gènes de pSymB est proche parente de gènes retrouvés dans le chromosome linéaire de *Agrobacterium tumefaciens*, incluant les gènes *rep* et *min*. Ceci signale une origine commune de ces réplicons. Des gènes dont le plus proche voisin appartient à la même espèce ont tendance à se regrouper en « amas ». L'ordre des gènes à l'intérieur de ces amas est conservé, mais le contenu des amas n'est pas limité qu'à des opérons. Nos données démontrent que 13 % des gènes de pSymB ont leurs plus proches voisins chez des espèces qui ne sont pas membres de la famille des *Rhizobiaceae* (incluant deux archées), et qu'ils représentent probablement des gènes impliqués dans un transfert horizontal.

**Mots clés:** *Sinorhizobium meliloti*, transfert horizontal, évolution de pSymB.

[Traduit par la Rédaction]

## Introduction

Five rhizobacterial genomes have been completely sequenced, including *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, *Mesorhizobium loti*, *Methylobacterium extorquens*, and *Brucella melitensis*. Several large plasmids belonging to other *Rhizobiaceae* have also been sequenced. The genome sequence data for these closely related species provides an opportunity for a thorough examination of genome organization and evolution of these species.

The organization of  $\alpha$ -proteobacterial genomes has been described as "unconventional" because members of this subgroup often have multiple chromosomes (occasionally linear) and

megabase-sized plasmids (Jumas-Bilak et al. 1998). *Sinorhizobium meliloti* is an endosymbiotic N<sub>2</sub>-fixing bacterium, whose genome is composed of a chromosome (3.7 Mb) and two large replicons, pSymA (1.4 Mb) and pSymB (1.7 Mb), with GC contents of 62.7, 62.4, and 60.4%, respectively (Galibert et al. 2001). The completed genome sequence of *S. meliloti* has shown that in addition to having a similar G+C content to the chromosome, the pSymB replicon carries essential genes (Finan et al. 2001), and comparative dinucleotide analysis has also revealed chromosome-like compositional properties of this replicon (Wong et al. 2002). These findings have led to the suggestion that the pSymB replicon should be designated a second chromosome in *S. meliloti*.

*Agrobacterium tumefaciens* is a plant pathogen causing crown gall disease. Its 5.67-Mb genome (strain C58) consists of one circular chromosome, one linear chromosome, and the plasmids pAt and pTi (Allardet-Servent et al. 1993). Comparative analysis using the National Center for Biotechnology Information's (NCBI) COG analysis and BLASTP identified 67% of the *A. tumefaciens* circular chromosomal genes as likely orthologs of *S. meliloti* chromosomal genes (Wood et al. 2001)

Received 5 November 2002. Revision received 25 April 2003. Accepted 1 May 2003. Published on the NRC Research Press Web site at <http://cjm.nrc.ca/> on 10 June 2003.

K. Wong and G.B. Golding,<sup>1</sup> McMaster University, Department of Biology, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada.

<sup>1</sup>Corresponding author (e-mail: [golding@mcmaster.ca](mailto:golding@mcmaster.ca)).

and showed that global gene order is highly conserved between these two replicons (Goodner et al. 2001). These observations support the idea that these chromosomes were derived from a recent ancestral chromosome (Wood et al. 2001). A strong syntenic relationship between pSymB and any of the *A. tumefaciens* replicons, however, was not observed, so the origin and evolutionary history of pSymB remain obscure.

The amount of horizontal transfer and the extent of rearrangement of the pSymB replicon is unknown. There is evidence (i) for horizontal transfer of insertion sequence (IS) elements from *S. meliloti* to *A. tumefaciens* (Deng et al. 1995), for (ii) horizontal transfer of glutamine synthetase between rhizobia (Turner and Young 2000), and (iii) that the *nod* genes, key genes involved in symbiosis, are thought to have spread among rhizobia by horizontal gene transfer (Suominen et al. 2001; Young and Johnston 1989). IS elements are found on pSymB, in addition to *Rhizobium*-specific intergenic mosaic elements (RIMES) and A, B, C palindromic elements (Galibert et al. 2001; Finan et al. 2001; Østeras et al. 1995, 1998), which may contribute to recombinations, rearrangements, and (or) horizontal transfer of pSymB genes. However, the G+C content and signature dinucleotide frequencies have been maintained at relative chromosome-like levels, suggesting that the level of recent horizontal transfer is low.

Although COG analysis can identify likely orthologs (Tatusov et al. 2000) of pSymB gene products, whole-genome nearest phylogenetic neighbor analysis allows for an investigation of evolutionary relationships to the closest relatives and for possible identification of horizontally transferred genes. This approach to whole-genome analysis, termed "phylogenomics" (Eisen 1998), has been applied to predict protein function. Sicheritz-Ponten and Andersson (2001) have used phylogenomics to analyze the evolution of microbial proteomes, focusing on biochemical pathways and horizontal gene transfer. Whole-genome phylogenetic analysis may help to decipher the origin of pSymB and how it has evolved to its present-day form, in terms of the degree of horizontal gene transfer and rearrangements.

A potential limitation to a phylogenetic approach to whole-genome analysis is the lack of similar sequences to a particular protein, in which case the nearest phylogenetic neighbor cannot be determined conventionally. Contextual information, such as conservation of gene order across taxa, can be used to predict operons (Ermolaeva et al. 2001) as well as gene function and functional interactions between genes (Huynen et al. 2000). Here, we explored the application of contextual information to predict the nearest neighbor in cases where phylogenetic trees cannot be constructed or when the assignment of a nearest neighbor is not statistically supported by bootstrap values.

In addition to using phylogenetic analysis, native and foreign genes may be identified by estimating the amount of time a gene has been residing in a genome. Native genes generally have a base composition and codon usage that is characteristic of the entire genome (Médigue et al. 1991). Through the process of amelioration, horizontally transferred genes acquire the base composition and codon usage of the recipient genome, and therefore, the time required for amelioration can be used as an estimate of the time of introgression (Lawrence and Ochman 1997). Here, the origin and evolution of the *S. meliloti* pSymB replicon was investigated using whole-genome nearest-

neighbor analysis and amelioration times. These methods reveal a complicated evolutionary history for this replicon.

## Methods

### Sequence data and determination of nearest phylogenetic neighbors

The complete *S. meliloti* genome, including all protein and nucleotide sequences for protein-coding regions, were downloaded from the sequencing consortium website at <http://sequence.toulouse.inra.fr/rhime/Complete/doc/Complete.html>. Each of the 1570 pSymB gene products were analysed. Initially, similarity searches were conducted using NCBI's BLASTP algorithm (August 23, 2001). A maximum of 50 hits (excluding any hits to other *S. meliloti* sequences) with an Expect (E) value less than  $10^{-20}$  were used in the phylogenetic analysis. Those pSymB proteins with less than three significant hits were excluded from further analysis. Sequence similarity searches to *S. meliloti* sequences were performed using the stand-alone BLAST search against a local database containing only *S. meliloti* protein sequences. Hits with E values less than  $10^{-20}$  were combined with the hits from the nonredundant NCBI protein database. The protein sequences were aligned using Clustal W (Thompson et al. 1994). The data was bootstrapped 100 times using SEQBOOT (PHYLIP version 3.5c; Felsenstein 1993), and distance matrices were generated with TREE-PUZZLE (version 5.0; Strimmer and von Haeseler 1996) using the JTT model of substitution and a Gamma model of rate heterogeneity (with eight rate categories).

Nearest phylogenetic neighbors were determined using the neighbor-joining method (NEIGHBOR program from PHYLIP version 3.5c). The protein separated from the pSymB protein by the fewest number of nodes was defined as the nearest phylogenetic neighbor. If there was more than one possibility, then the nearest neighbor was chosen by determining the minimum total branch length from each of the possible nearest neighbors. The nearest neighbor was assigned if the corresponding bootstrap value was 95% or greater. If this was not the case, contextual information was used to help predict the nearest neighbor. Initially, the neighbor with the highest bootstrap value was chosen as the potential nearest neighbor. One of the following two conditions were satisfied for a nearest-neighbor assignment to be made.

1. The gene is flanked on both sides by genes that have a nearest neighbor in the same species, with bootstrap values  $\geq 95\%$ ;
2. On one side only, the gene is next to at least two successive genes that have a nearest neighbor in the same species, with bootstrap values  $\geq 95\%$ .

Nearest-neighbor assignments were also made for proteins for which a tree could not be constructed because of the lack of sequences having BLAST E values less than  $10^{-20}$  (see above). In this case, a prediction was made only if condition 1 was satisfied.

### Identification of potential operons

An operon can be described as a group of genes that are transcribed into a single mRNA molecule. We followed the method of Ermolaeva et al. (2001) on prediction of operons and

their observation that operon gene order and orientation tend to be conserved across genomes. Because we did not identify potential operons experimentally, they were identified based on the following criteria: a potential operon is a group of two or more successive genes with the same nearest neighbor, coded on the same strand with intergenic regions less than or equal to 200 bp.

### Dinucleotide analysis

The overall dinucleotide signature of pSymb was determined using methods previously described (Karlín and Burge 1995). For analysis of regions within pSymb, dinucleotide relative abundances were determined for 50-kb windows, with a 10-kb overlap, and  $\delta^*$ -distances were determined between each 50-kb region and the overall dinucleotide signature.

### Calculation of substitution rates for *S. meliloti*

Glutamine synthetase I and glutamine synthetase II have been shown to behave as good molecular clocks (Pesole et al. 1991). Using glutamine synthetase I and glutamine synthetase II sequences, Turner and Young (2000) have estimated divergence times among Rhizobia. They made four estimates for the *S. meliloti* – *M. loti* split, and the average value of 265 million years was used here to calculate substitution rates.

Substitution rates were determined by comparing 30 pairs of highly conserved nucleotide sequences from *S. meliloti* and *M. loti*: 25 ribosomal protein genes and five elongation factors (*tufA*, *tufB*, *sigA*, *fusA*, and *infA*). Synonymous and nonsynonymous substitution rates ( $d_S$  and  $d_N$ ) were calculated using codeml in the PAML program package (version 3.1; Yang 1997), which applies the codon substitution model of Goldman and Yang (1994). The transition to transversion ratio was estimated by the algorithm, and a global clock was assumed. Each of the 30 pairs have  $d_S$  less than 1.5. The mean  $d_S$  was 1.16 and the mean  $d_N$  was 0.08. Based on a divergence time of 265 million years, the following substitution rates were calculated for the first, second, and third codon positions, respectively: 0.0262, 0.0160, and 0.1626% substitutions per million years per lineage. The mean transition to transversion ratio was estimated to be 1.6, and this value was used for the amelioration simulation.

### Amelioration simulation

Muto and Osawa (1987) observed a linear correlation between the G+C content of a genome and the G+C content of each codon position in coding regions. Lawrence and Ochman (1997) describe these relationships as

$$[1] \quad GC_{1st} = 0.615 \times GC_{Genome} + 26.9$$

$$[2] \quad GC_{2st} = 0.270 \times GC_{Genome} + 26.7$$

$$[3] \quad GC_{3st} = 1.692 \times GC_{Genome} - 32.3$$

where  $GC_{Genome}$  is the G+C content of the total genome. Lawrence and Ochman (1997) validated this model for  $20\% \leq GC_{Genome} \leq 80\%$ . They also determined that the use of fewer than 1500 codons gave unreliable estimates of codon-specific G+C content and inaccurate estimates of the amelioration time. Therefore, pSymb genes were pooled according to similarity of G+C contents in the first, second, and third codon positions,

with no fewer than 1500 codons (4.5 kb) per group. From these pooled genes, the G+C content of each codon position was determined, and the amelioration time was estimated for the entire pool.

We conducted simulations starting from random sequences generated with codon-position-specific G+C contents for each  $GC_{Genome}$  ranging from 20 to 80%. To simulate the mutational biases of the *S. meliloti* genome, Tamura's (1992) model of substitution was applied to each codon position, and each sequence was ameliorated in million-year (Myr) intervals for 1000 Myr toward the average codon-position-specific G+C contents of pSymb genes: 64.8, 46.3, and 76.8%, respectively, for the first, second, and third codon positions. This was iterated 1000 times for each  $GC_{Genome}$ .

To estimate the amelioration time for each pool of genes, the G+C contents of the generated sequences were compared with the actual sequences. For each Myr interval, the base composition of the generated sequence was determined, and the weighted least-squares difference was calculated between the G+C contents of each codon position in the generated sequence and the pooled pSymb genes. Because the third codon position ameliorates more quickly, while the first and second ameliorate more slowly, the weights 3.7, 1.3, and 5.0 were applied to the first, second, and third codon positions, respectively (Lawrence 1995). The mean least-squares difference was calculated for the 1000 iterates at each Myr interval. The overall minimum least-squares difference was determined, and the time at which this minimum occurred was taken as the best estimate of the time of introgression of the pooled pSymb genes. The initial  $GC_{Genome}$  that gave the minimum least-squares difference was taken as the G+C content of the genes at the time of introgression.

## Results

### Nearest-neighbor analysis

Although the BLAST algorithm is often used to determine the similarity of a gene or protein to those in databases, it has been shown that the best BLAST hit is often not the nearest phylogenetic neighbor (Koski and Golding 2001). Therefore, the closest relative of *S. meliloti* pSymb protein sequences were determined using a phylogenetic method. Table 1 summarizes the nearest-neighbor analysis.

In total, 510 nearest-neighbor assignments were made. This represents approximately one-third of pSymb ORFs. Contextual information allowed us to assign nearest neighbors to 31 genes with nearest-neighbor bootstrap values under 95%. For the 19 nearest-neighbor predictions based on contextual information and no phylogenetic information (because of limited sequence data), the BLAST output for those genes were inspected to confirm that a homologous protein in that species exists. Only one gene did not have a homolog in the predicted species, indicating that when combined with phylogenetic information, contextual information is a reliable method to predict the nearest neighbor of a protein.

### Approximately 13% of pSymb genes have been involved in horizontal gene transfer

Of the 33 species with nearest neighbors to pSymb genes, 31 are *Bacteria*, and two are *Archaea* belonging to the *Eur-*

**Table 1.** Summary of nearest-neighbor analysis of pSymB ORFs.

	No. of genes	% of total
Total protein-coding genes	1570	100
One or more BLAST hits* with Expect values $\leq 10^{-20}$	1211	77
Phylogenies constructed <sup>†</sup>	984	63
Nearest-neighbors with significant bootstraps <sup>‡</sup>	460	29
Predictions for ORFs with nonsignificant bootstraps <sup>§</sup>	31	2.0
Predictions for ORFs with no phylogeny constructed <sup>§</sup>	20	1.3
Confirmed false predictions	1	0.06
<b>Total nearest neighbors predicted using contextual information</b>	<b>50</b>	<b>3</b>
<b>Total nearest-neighbor assignments</b>	<b>510</b>	<b>33</b>

\*BLAST hits to genes in species other than *Sinorhizobium meliloti*.

<sup>†</sup>Phylogenies were constructed for pSymB ORFs with three or more BLAST hits.

<sup>‡</sup>If the same nearest neighbor occurred in at least 95 out of 100 bootstrapped trees for a particular ORF, it was assigned as the nearest neighbor.

<sup>§</sup>For ORFs without significant nearest-neighbor bootstraps or no phylogeny constructed, assignments were made using contextual information where possible (see Methods).

*yarchaeota* phylum (Table 2). Nearly half (45.1%) of the pSymB proteins are nearest neighbors to *A. tumefaciens* proteins. The majority of these *A. tumefaciens* proteins are encoded on the linear chromosome (140 out of 230) while the remainder are encoded on the circular chromosome (66 out of 230) and pAt (24 out of 230). None are found on the pTi plasmid, despite the COG analysis indicating that many likely orthologs of pSymB proteins are present (Wood et al. 2001). *Mesorhizobium loti* proteins are the nearest neighbors for approximately 28% of the pSymB gene products, and nearly 11% are related to other *S. meliloti* proteins, which likely are results of gene duplications (Table 2). Another 3% have nearest neighbors from other members of the *Rhizobiaceae* family. These data suggest that *A. tumefaciens*, *M. loti*, and *S. meliloti* form a clade of closely related species, despite their divergence into pathogenic and symbiotic lifestyles. In total, 87% of pSymB genes are a result of a duplication or have a nearest neighbor in other *Rhizo*biaceae.

Approximately 10% of pSymB genes have nonrhizobial nearest neighbors in the  $\alpha$ -,  $\beta$ -, and  $\gamma$ -proteobacterial subgroups (Table 2). In many instances, a similar protein is absent in *A. tumefaciens* and (or) *M. loti*, or other  $\alpha$ -proteobacteria for which sequences are available. Alternatively, the nonrhizobial nearest neighbor is clustered within a clade that includes *S. meliloti*, *A. tumefaciens*, and *M. loti* proteins. The remaining 3% of the pSymB proteins have nearest neighbors from the distantly related *Deinococcus* group, cyanobacteria (blue-green algae), six species of Gram-positive bacteria, and two archaea (Table 2). The 13% of pSymB genes with these unusual phylogenetic relationships have likely been involved in horizontal transfer. (The complete list of these genes can be accessed at <http://life.biology.mcmaster.ca/~brian/pSymB-HTgenes.html>) Although the list of nearest neighbors represents a wide spectrum of bacterial species, all are pathogens and (or) inhabitants of soils, sediments, or water; their shared environment with *S. meliloti* may provide opportunities for genetic exchanges. There does not appear to be a correlation between the location of these genes and the location of IS elements; RIMEs; and A, B, C palindromic elements (Fig. 1). Almost half (41.5%) of the pSymB genes involved in horizontal transfer are involved in the

metabolism of small molecules.

The pSymB replicon is known to play an important role in the early stages of nodule formation and symbiotic nitrogen fixation. Our analysis shows that some of these genes have been involved in horizontal transfer. Only four proteins coded on pSymB have been identified as nodulation proteins, NodP, NodQ, NfeD, SMB20472, a putative nodulation protein belonging to the NodU family. Nearest-neighbor analysis confirms previous suggestions that the *nodPQ* genes have been duplicated on pSymA (Galibert et al. 2001), and that a protein from *Rhizobium etli* is the nearest neighbor to NfeD, a protein involved in nodulation competitiveness. SMB20472 has likely been involved in horizontal transfer, since the nearest neighbor, with which it shares 62% identity, is a transferase from the cyanobacterium *Synechocystis* sp. PCC 6803. Genes involved in surface polysaccharide biosynthesis are important for successful nodule invasion (Hynes et al. 1986). In our analysis, a nearest neighbor was not assigned to many of these genes; however, we were able to identify some that were potentially involved in horizontal transfer. Nearest neighbors to these genes are found in *Caulobacter vibrioides* and *Pseudomonas aeruginosa*.

One of two  $\beta$ -galactosidase genes and a gene coding for a putative membrane protein (SMB20185) have nearest neighbors found in an archaeal species. Analysis of these phylogenies suggests that both of these genes were horizontally transferred into the *S. meliloti* genome from an archaea. The only two  $\beta$ -galactosidase genes in *S. meliloti* are coded on pSymB. Phylogenetic analysis supports that the direction of transfer for each of these genes, from a distantly related donor, is into the *S. meliloti* genome. One copy (*lacZ1*) appears to have been acquired from *Thermoanaerobacterium thermosulfurigenes* and the other (*lacZ2*) from the halophilic archaea *Haloferax alicantei*. Unlike the *lacZ1* gene, the *lacZ2* gene is not clustered on pSymB with lactose transporter genes and the LacZ2 protein has significant similarity only to archaeal  $\beta$ -galactosidases.

A putative membrane protein (GenBank accession No. AAB85581.1) from the archaea *Methanothermobacter thermoautotrophicus* has been determined as the nearest neighbor of the putative solute-binding membrane protein encoded by

**Table 2.** *Sinorhizobium meliloti* pSymB nearest neighbors.

Species	No. of nearest neighbors	Prediction <sup>†</sup>	Proportion <sup>‡</sup> (%)
<i>α-Proteobacteria</i>			
<i>Sinorhizobium meliloti</i> *	50	4	10.6
<i>Agrobacterium tumefaciens</i> *	207	23	45.1
<i>Mesorhizobium loti</i> *	127	17	28.2
<i>Rhizobium leguminosarum</i>	4		0.8
<i>Sinorhizobium fredii</i>	4	1	1.0
<i>Rhizobium etli</i>	3		0.6
<i>Rhizobium</i> sp. NGR234	2		0.4
<i>Brucella melitensis</i>	1		0.2
<i>Methylobacterium extorquens</i> *	1		0.2
<i>Non-Rhizobiaceae</i>			
<i>Caulobacter vibrioides (crescentus)</i>	6		1.2
<i>Paracoccus denitrificans</i>	4		0.8
<i>Paracoccus pantotrophus</i>	1		0.2
<i>Rhodobacter sphaeroides</i>	3	1	0.6
<i>Rhodobacter capsulatus</i>	1		0.2
<i>β-Proteobacteria</i>			
<i>Bordetella bronchiseptica</i>	1		0.2
<i>Ralstonia</i> sp.	1		0.2
<i>γ-Proteobacteria</i>			
<i>Pseudomonas aeruginosa</i> *	17	2	3.7
<i>Vibrio cholerae</i> *	5	1	1.2
<i>Pasteurella multocida</i> *	4	1	1.0
<i>Pseudomonas syringae</i>	1		0.2
<i>Acidithiobacillus ferrooxidans</i>	1		0.2
<i>Escherichia coli</i> *	1		0.2
<i>Yersinia pseudotuberculosis (pestis)</i> *	1		0.2
High G+C, Gram <sup>+</sup>			
<i>Streptomyces coelicolor</i>	3		0.6
<i>Mycobacterium tuberculosis</i> *	2		0.4
<i>Mycobacterium avium</i>	1		0.2
<i>Streptomyces hygroscopicus</i>	1		0.2
Low G+C, Gram <sup>+</sup>			
<i>Desulfonispora thiosulfatigenes</i>	1		0.2
<i>Thermoanaerobacterium</i> sp.	1		0.2
Thermus–Deinococcus			
<i>Deinococcus radiodurans</i> *	2		0.4
Cyanobacteria			
<i>Synechocystis</i> sp.*	1		0.2
Euryarchaeota			
<i>Haloferax alicantei</i>	1		0.2
<i>Methanothermobacter thermoautotrophicus</i> *	1		0.2
<b>Total</b>	<b>460</b>	<b>50</b>	<b>100</b>

**Note:** The total number of pSymB ORFs is 1570.

\*Completely sequenced genomes at the time the BLAST search was performed.

<sup>†</sup>Nearest neighbors assigned based on contextual information.

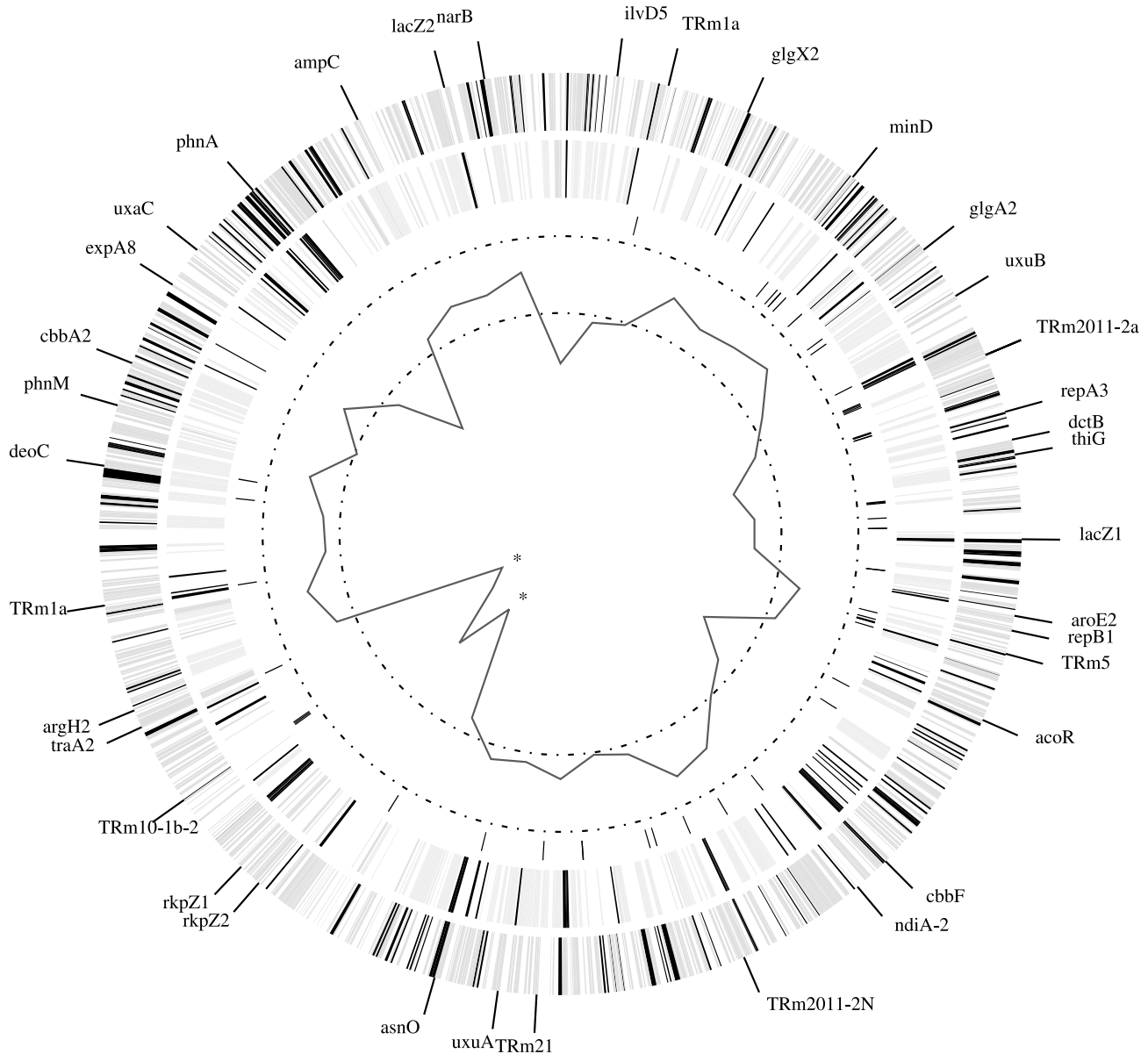
<sup>‡</sup>The percentage of nearest neighbors assigned out of the total.

SMb20185. This protein is part of a putative ABC-type transporter on pSymB. The SMb20185 protein has significant sequence similarity to other archaeal membrane proteins from *Pyrococcus* spp., *Archaeoglobus fulgidus*, *Thermoplasma volcanium*, *Sulfolobus solfataricus*, and *Aeropyrum pernix*, as well as the bacterium *Thermotoga maritima*.

#### Local gene order is conserved in “patches” that likely contain at least one operon

Here we define a “patch” as two or more successive genes with a nearest neighbor from the same species. A patch may or may not be composed of one or more potential operons. Two-thirds of the genes with assigned nearest neighbors are arranged

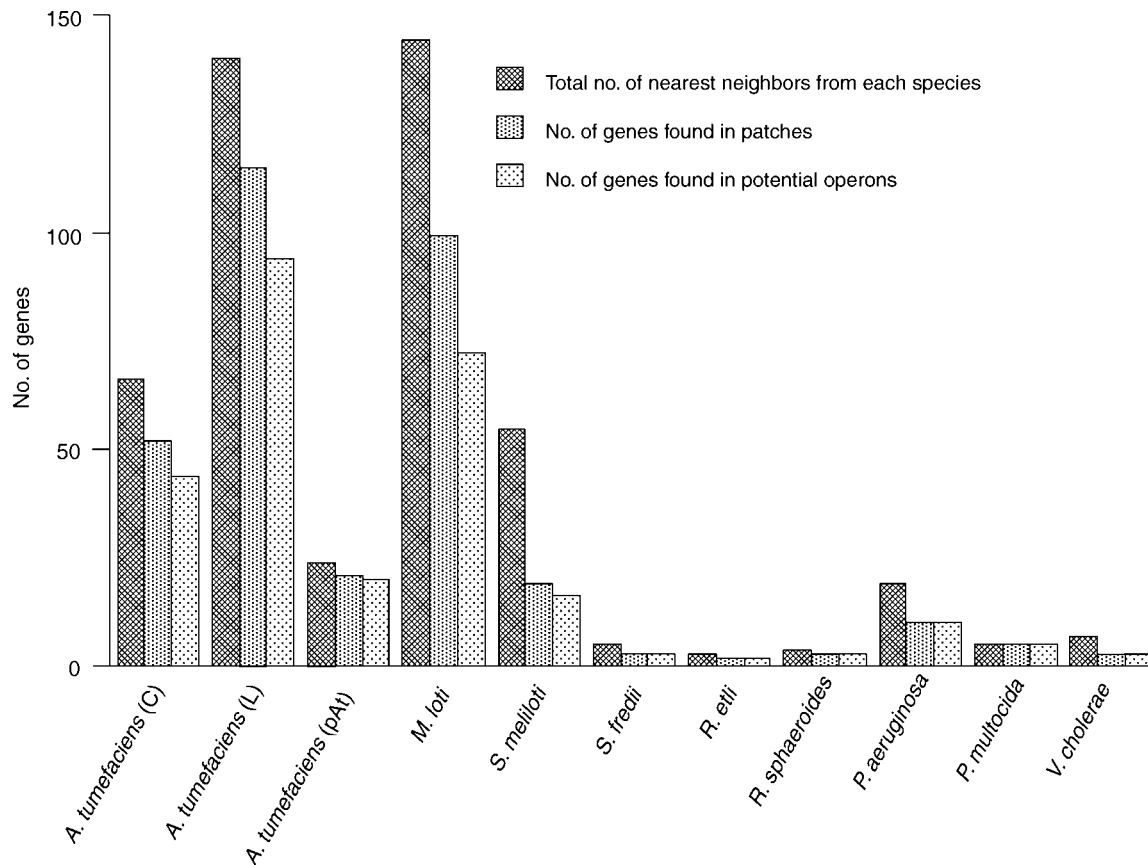
**Fig. 1.** Amelioration of pSymB genes and genes potentially involved in horizontal transfer. Shown on the outer circle are the genes for which simulation of amelioration was successful (light grey). Genes that are estimated to have been ameliorating for less than 10 million years are shown in black. The second circle from the outside shows the distribution of genes with nonrhizobial nearest neighbors (black) and rhizobial nearest neighbors (light grey). The third circle shows the location of IS elements; RIMEs; A, B, and C palindromic elements; and short partial IS elements and group II introns (black). The solid line shows the  $\delta^*$ -distance of overlapping 50-kb regions from overall (outer broken line) and the average (inner broken line) dinucleotide abundances (see Methods). Asterisks (\*) indicate regions with the largest  $\delta^*$ -distances.



in patches. The distribution of these patches among the different nearest neighbors is shown in Fig. 2. Most (187 out of 230 or 81%) of the pSymB genes encoding proteins with a nearest neighbor from *A. tumefaciens* are grouped in patches that range from two to nine genes. Comparison of gene order between *A. tumefaciens* and *S. meliloti* patches shows that there have been some local rearrangements since their divergence; however, gene order is conserved in most of these patches. Nearly two-thirds of these genes have nearest neighbors located on the *A. tumefaciens* linear chromosome, suggesting a common origin between the two replicons.

Nearest neighbors to *M. loti* genes are also arranged in patches, although to a lesser extent. Patches comprise 69% (99 out of 144) of the genes with *M. loti* nearest neighbors (Fig. 2) and they consist of 2–11 genes. Gene order is not well conserved. The exception is one patch of 11 genes (SMb20422 to SMb20432) located adjacent to the IS element ISRM21, in which gene order is completely conserved. Six patches comprise 35% of the genes that were determined to have *S. meliloti* nearest neighbors. Gene order is conserved in most of these patches, which implies that the duplication of these genes occurred as a single event rather than several independent events.

**Fig. 2.** Distribution of patches and potential operons among species with nearest neighbors to pSymB genes. C, L, and pAt refer to the *Agrobacterium tumefaciens* circular and linear chromosomes and plasmid pAt, respectively.



Six other species are represented in the remaining patches, three  $\alpha$ -Proteobacteria (*Sinorhizobium fredii*, *Rhizobium etli*, *Rhodobacter sphaeroides*), and three  $\gamma$ -Proteobacteria (*Pseudomonas aeruginosa*, *Vibrio cholerae*, *Pasteurella multocida*). On average, these patches consist of fewer genes than the *A. tumefaciens* patches, ranging from two to five genes per patch. Comparison of gene order between *S. meliloti* and these species showed that a rearrangement or an insertion and (or) deletion of a gene has occurred in only two of these patches.

Potential operons were identified within patches, and occasionally these operons were flanked by one or more genes that were in the same patch but not included in the operon. Such cases are only observed for patches with nearest neighbors to *A. tumefaciens*, *M. loti*, and *S. meliloti* genes (Fig. 2). Overall, few patches consist of two or more potential operons (9 out of 86) or no potential operons (12 out of 86).

As stated above, pSymB genes with nearest neighbors from species outside of the *Rhizobiaceae* family are likely to have been involved in horizontal transfer. The patches with nearest neighbors from *R. sphaeroides*, *P. aeruginosa*, *V. cholerae*, and *P. multocida* are all composed of one (or two) potential operons with no adjacent "single" genes. This and conservation of gene order indicate that genes that were horizontally transferred together likely comprise an operon. However, two-thirds of the genes involved in horizontal transfer are not part of patches or operons, and in many cases this is because only one nearest neighbor from a particular species has been identified (Table 2).

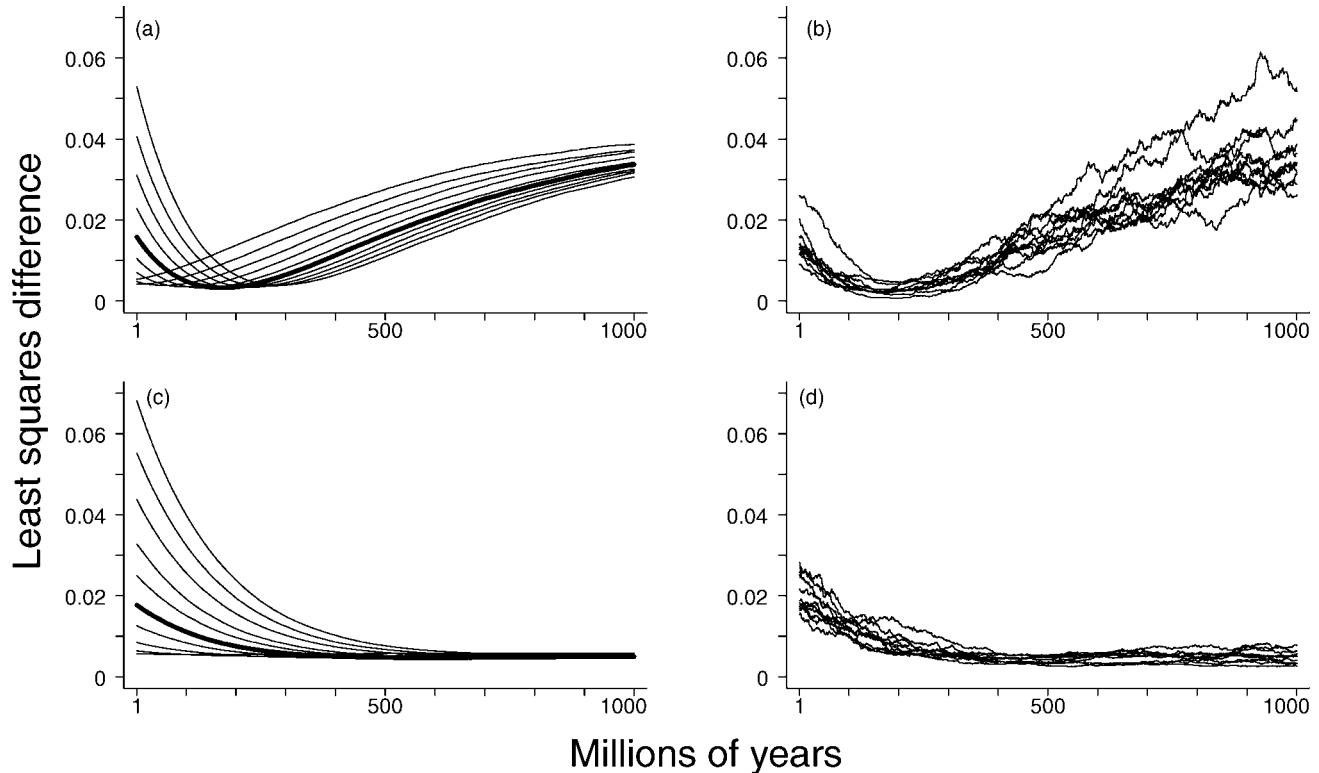
These observations show that although horizontal transfer of operons may occur, the long term result of horizontal transfer tends to predominantly result in transfer of single genes.

### Simulation of amelioration

For all 1570 protein-coding genes on pSymB, a computer simulation of amelioration was applied to determine the initial G+C content and the time required for a sequence to reach its current codon-position-specific G+C contents. The estimation was considered valid only if a minimum least-squares difference in the G+C contents (of all three codon positions) between the simulated data and the actual data was reached at a time between 0 and 1000 Myr. These genes are shown in light grey in the outer circle of Fig. 1 or in black if the time was less than 10 Myr. The 725 genes for which a minimum between 0 and 1000 Myr was not reached were excluded from further analysis.

The average least-squares difference of 1000 iterates at each Myr interval was used to produce amelioration curves. Figure 3a shows the amelioration curves for 4.5-kb sequences with initial G+C contents of 49–58%, with the least-squares differences calculated between the generated sequences and a pool of genes with the present-day G+C contents of 58.8, 47.1, and 64.5% at the first, second, and third codon positions, respectively. The simulation gave an estimate of 172 Myr as the amelioration time, with the sequence having a G+C content of 53% at the time of introgression. According to the Muto and Osawa (1987) relationships, the G+C contents at the first, second, third codon

**Fig. 3.** Amelioration simulation. (a) Amelioration curves (mean of 1000 iterates) for sequences with initial G+C contents of 49–58%. The minimum least-squares difference was reached after 172 Myr, when the initial G+C content was 53% (bold line). (b) Ten iterates of amelioration of a 4.5-kb sequence with an initial G+C content of 53%. The G+C contents of sequences in (a) and (b) were compared with a pool of genes with 58.8, 47.1, and 64.5% G+C at the first, second, and third codon positions, respectively. (c) Amelioration curves for sequences with initial G+C contents of 54–65%. The minimum least-squares difference was reached after 586 Myr when the initial G+C content was 59% (bold line). (d) Ten iterates of amelioration of a 4.5-kb sequence with an initial G+C of 59%. The G+C contents of sequences in (c) and (d) were compared with a pool of genes with 62.4, 50.6, and 74.7% G+C at the first, second, and third codon positions, respectively.



positions were 59.4, 41.0, and 57.3%, respectively. In Fig. 3a, the largest and smallest minima differ by less than 0.002, yet the time required to reach these minima ranged from 1 to 287 Myr. At 172 Myr, the minimum least-squares difference was 0.00322, with a standard deviation of 0.00130. Two-sample *t* tests showed that this minimum is not significantly different than that reached by sequences with initial G+C contents of 52 and 54% (at a 5% level), giving a range of 135–204 Myr for the amelioration time. Figure 3b shows ten iterates of the simulation for sequences with an initial G+C content of 53%. The simulations demonstrate the large variance in the estimates of amelioration time. Each individual realization of the evolutionary history shows a completely different level of G+C content over time.

Figure 3c shows the amelioration curves for sequences with initial G+C contents of 54–65%. The pool of genes to which these sequences were being compared had G+C contents of 62.4, 50.6, and 74.7% at the first, second, and third codon positions, respectively. The minimum least-squares difference, averaged over 1000 iterates, occurred after 586 Myr with a sequence with an initial G+C content of 59% (63.1, 42.6, and 67.5% at the first, second, and third codon positions, respectively). However, the curves do not reach a minimum value. The apparent amelioration occurred by chance because of the high variation

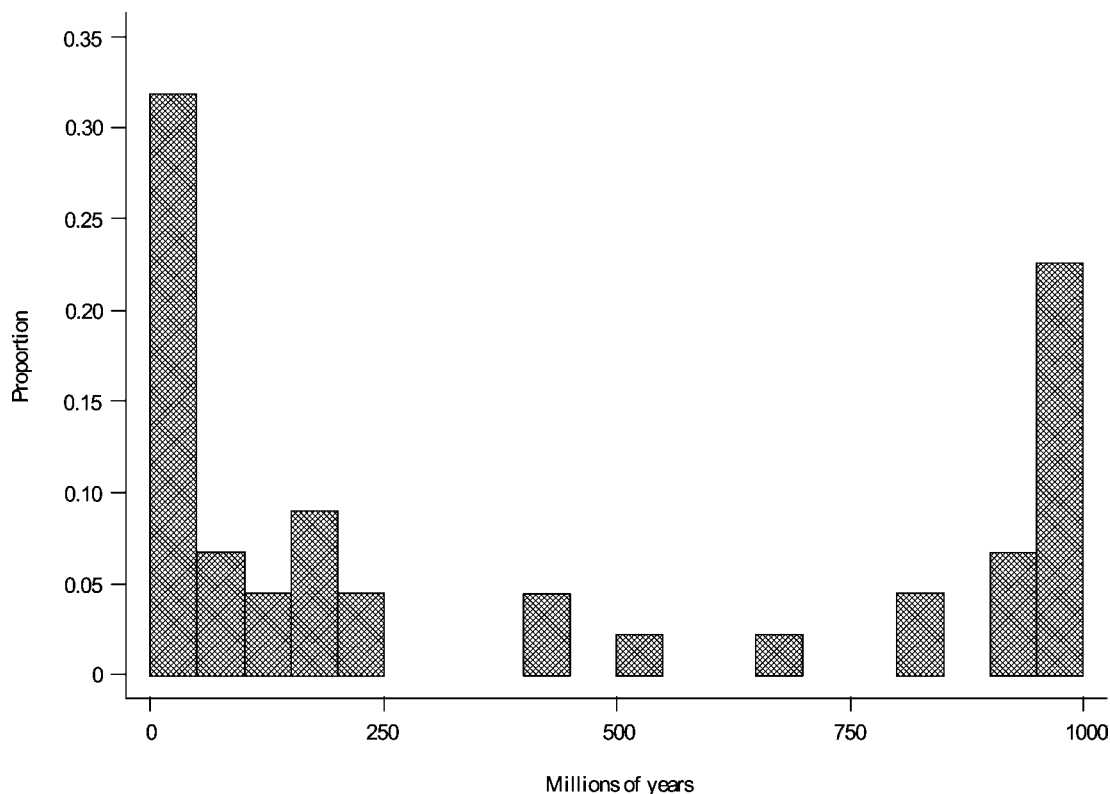
of the least-squares difference. This variation is demonstrated in Fig. 3d (as well as Fig. 3b), which shows ten iterates of the simulation involving sequences with an initial G+C content of 59%. Since only one of these realizations might have occurred, the estimation of amelioration times may be highly variable and unreliable.

However, amelioration times can provide gross estimates of potential times of introgression for some genes. Figure 4 shows amelioration times for 845 pSymB genes. Genes that are predicted to have been acquired by pSymB less than 10 Myr ago are distributed throughout the replicon in no apparent pattern (Fig. 1). This group of genes comprises 148 of the 845 genes that were assigned an amelioration time, and the species to which the nearest neighbor belongs is shown in Table 3. The majority of these genes (61%) do not have a nearest neighbor assigned, and proteins found in *Rhizobiaceae* make up 30% of this group. The remaining genes have distantly related nearest neighbors, and the predicted amelioration times of less than 10 Myr suggest that these genes were horizontally transferred recently into the *S. meliloti* genome.

#### Dinucleotide analysis

It has been shown that dinucleotide biases are consistent throughout a genome and distinguishable between species

**Fig. 4.** Estimated amelioration times of pSymB genes. Simulations that gave an estimate of 0 or 1000 million years are excluded. Proportions are calculated from a total of 845 genes. The initial G+C content of these genes ranges from 49 to 79%, with a mean of 64%.



**Table 3.** Species with nearest neighbors to pSymB genes that are estimated to have been ameliorating for less than 10 million years.

Nearest neighbor	No. of nearest neighbors	Total nearest neighbors*
<i>Sinorhizobium meliloti</i>	1	54
<i>Mesorhizobium loti</i>	19	147
<i>Agrobacterium tumefaciens</i>	23	230
<i>Rhizobium leguminosarum</i>	1	4
<i>Brucella melitensis</i>	1	1
<i>Pseudomonas aeruginosa</i>	6	19
<i>Vibrio cholerae</i>	2	6
<i>Paracoccus denitrificans</i>	1	4
<i>Thermoanaerobacterium</i> sp.	1	1
<i>Desulfonisporea thiosulfatigenes</i>	1	1
<i>Streptomyces hygrosopicus</i>	1	1
<b>None assigned</b>	<b>91</b>	<b>1060</b>

\*The total number of nearest neighbors assigned from a species.

(Burge et al. 1992; Karlin et al. 1994, 1997). A measure of dinucleotide bias for all 16 dinucleotides is the “dinucleotide relative abundance profile”, the frequency of a dinucleotide relative to expected values based on nucleotide content (Burge et al. 1992). Since genomic profiles are often unique to a species, this has been termed the “genome signature”. Dinucleotide analy-

sis of overlapping 50-kb regions of pSymB revealed regions with large differences in dinucleotide abundances, relative to the overall abundances and the average abundances in the 50-kb regions (Fig. 1). The two regions with the largest differences from the overall signature (indicated with asterisks in Fig. 1) contain many small unknown or hypothetical ORFs, transposons, and genes involved in sugar-nucleotide biosynthesis and exopolysaccharide synthesis. In addition, the G+C content of this region is slightly below 60%. The unusual dinucleotide composition and G+C content of this large region suggest it may have a foreign origin. However, many of the duplicated pSymB genes are found in this region (data not shown), suggesting that the small, hypothetical ORFs may represent degenerating genes that may have been duplicated but not maintained on pSymB. Regions with the smallest differences include those with the *repA1B1C1* genes, *nodP2Q2*, and a large cluster of *exp* genes.

## Discussion

Sequencing of the complete *S. meliloti* genome has revealed the mosaic nature of this genome. This is demonstrated in the (i) functions of the genes distributed among the three replicons; (ii) G+C content and codon usage; and (iii) numbers and distribution of IS elements, RIMEs, and A, B, C palindromic elements (Galibert et al. 2001). It has been argued that the pSymB replicon is a second chromosome in *S. meliloti* (Finan et al. 2001). Unlike the main chromosome, the pSymB replicon does not

show a high degree of synteny with an *A. tumefaciens* replicon, and orthologs are found on all four *A. tumefaciens* replicons (Goodner et al. 2001; Wood et al. 2001). Therefore, the origin and evolutionary history of pSymB are less obvious.

Several aspects were examined in an attempt to determine how pSymB might have evolved. Dinucleotide analysis shows regions with biases that deviate from the average abundances on pSymB, but since this analysis is limited to regions 50 kb or larger, methods that allow a finer gene-by-gene examination are required. A phylogenomic approach (Eisen 1998), whole-genome nearest-neighbor analysis, provides evolutionary information and enables the origins of genes and groups of genes to be examined. In addition to determining the most closely related genes, phylogenetic relationships indicate which genes may have been involved in horizontal gene transfer. The length of time a gene has spent in a genome can also point to horizontal transfer, and it has been argued that amelioration time corresponds to the time of introgression (Lawrence and Ochman 1997).

A nearest neighbor was assigned for one-third of the pSymB protein-coding genes. In addition to assignments of the nearest neighbor based on statistically significant data, we made additional assignments based on contextual information. Contextual information has also been used to predict protein function and functional interaction (Huynen et al. 2000); these predictions are also based on the observation that gene order is conserved in operons and across genomes. Here, it has been demonstrated that contextual information can supplement phylogenetic information.

These data show that the pSymB replicon is truly mosaic; duplicated genes and genes with nearest neighbors to 32 other species are dispersed seemingly randomly throughout pSymB, and gene order is conserved only locally in small groups of genes. Because of this random distribution and because the proportion of genes involved in horizontal transfer is small (13%), the dinucleotide signature remains similar to that of the main *S. meliloti* chromosome. The complexity of pSymB is also contrasted with the scarce number of IS elements and other repeat elements found on pSymB. An estimation of 13% of the pSymB genes, either acquired by horizontal transfer or horizontally transferred to other species, is comparable to other estimates of horizontally acquired genes in other species (Garcia-Vallve et al. 2000), which range from 1.5% to 14.5%. These estimates were determined using a statistical procedure that identified genes with unusual G+C content, codon usage, and amino acid usage; gene position was also taken into consideration. This is a more stringent procedure than that employed by Lawrence and Ochman (1998), who used atypical G+C content and codon bias to identify horizontally acquired genes in *Escherichia coli*. This method, however, may falsely identify native genes as horizontally acquired genes (Koski et al. 2001; Wang 2001). Lawrence and Ochman (1998) estimate that approximately 18% of *E. coli* genes have been acquired since its divergence from *Salmonella* 100 million years ago, while Garcia-Vallve et al. (2000) estimate this value to be 9.6%. This concurs with the estimate of 10–15%, based on phylogenetic analysis (Koski et al. 2001).

Although genes with nearest neighbors found in members of the *Rhizobiaceae* family were excluded here as potentially involved in horizontal transfer, there is strong evidence for horizontal transfer among rhizobia. It has been shown that an IS

element has been horizontally transferred from *S. meliloti* to *A. tumefaciens* (Deng et al. 1995), and the incongruence of phylogenies based on *nod* genes with those based on 16S rDNA shows that *nod* genes have been horizontally transferred among rhizobia (Suominen et al. 2001; Young and Johnston 1989). It has also been suggested that glutamine synthetase II may have been horizontally acquired by *Bradyrhizobium japonicum*, *Mesorhizobium huakuii*, and *Rhizobium galegae* from other rhizobia (Turner and Young 2000). Here, it was observed that the largest patch has *M. loti* genes as the nearest neighbors, and its location adjacent to a transposon indicates it was likely acquired intact from *M. loti*. In addition, the *A. tumefaciens* genes most closely related to pSymB genes are dispersed among three of the four *A. tumefaciens* replicons, and this is likely a result of horizontal transfer between these species as well as vertical transmission from a common ancestor. Our estimate of 13% of pSymB genes involved in horizontal transfer, therefore, may be a conservative estimate.

Genes involved in horizontal transfer that are found in patches comprise exactly one or two operons (Fig. 2) in which gene order is conserved. This shows that the genes within an operon were transferred as a single unit rather than in separate transfer events. Comparison of operon structures in complete microbial genomes has shown that generally, gene order within operons was found to be less conserved with longer-term evolution and higher numbers of IS elements (Itoh et al. 1999). Therefore, given the conservation of gene order within the horizontally transferred operons on pSymB, they may have been involved in recent horizontal transfers. However, it appears that in *S. meliloti*, horizontal transfer of single genes has occurred more often than transfer of operons, since two-thirds of the predicted horizontally transferred genes are not found in potential operons.

Nearest-neighbor analysis can indicate which genes may have a foreign origin but does not provide an estimate of the time of introgression. The amelioration simulation yielded estimates with high variation and did not give estimates of introgression times for some genes. For these genes the least-squares difference increased with time, resulting in a minimum least-squares difference at 0 Myr, or the least-squares difference decreased with time but did not reach a minimum even after 1000 million years of amelioration. Nor was there a strong agreement of amelioration times among genes within potential operons that have been involved in horizontal transfer (data not shown). Some problems with this method include inaccurate amelioration times for (i) genes that have been horizontally acquired from a donor with a similar genomic G+C contents at each codon position and (ii) genes with unusual codon usage or G+C contents because of functional constraints (Jukes and Kimura 1984; Miyata and Yasunaga 1980). Therefore, the amelioration time may not correlate with the actual time of a gene's residence in a genome, and estimates given by this method should not be used alone as an indicator of horizontal transfer.

There are conflicting views concerning the origins of the *S. meliloti* pSymA and pSymB replicons. Galibert et al. (2001) suggest that these replicons were acquired by an ancestral rhizobium and that pSymA was acquired more recently than pSymB because overall G+C content and codon usage are distinctive from pSymB and the chromosome. These authors argue against a chromosomal origin for pSymB, citing the small number of

IS elements and large proportion of unique genes. With the completion of the *A. tumefaciens* genome, BLAST and COG analyses revealed a substantial number of orthologs between all three *S. meliloti* replicons and all four *A. tumefaciens* replicons (Goodner et al. 2001; Wood et al. 2001). Wood et al. (2001) suggest that ancestral pSymA and pSymB replicons were acquired prior to divergence of *A. tumefaciens* and *S. meliloti*.

In addition to having a large portion of the nearest neighbors to pSymB genes, most of which are organized in operons where local gene order is conserved, key genes on the *A. tumefaciens* linear chromosome involved in plasmid replication and cell division share a common origin with those on pSymB. Together, these observations suggest that pSymB and the *A. tumefaciens* linear chromosome have a common origin, which was a plasmid present in the last common ancestor of *S. meliloti* and *A. tumefaciens*.

Both the linear chromosome and pSymB have a *repABC* system of replication, which is common among  $\alpha$ -proteobacterial plasmids (Tabata et al. 1989; Palmer et al. 2000). The *S. meliloti* *repA1B1* are involved in pSymB segregation, and *repC1* is involved in pSymB replication (Chain et al. 2000). The ability to self-replicate is an essential property of plasmids, and genes that confer this property must be present for a plasmid to survive (Thomas 2000). Therefore, the *rep* genes found on a replicon are likely also to have been present in its ancestral plasmid. Our analysis showed that indeed, the pSymB RepA1B1 proteins are nearest neighbors to the *repAB* gene products from the *A. tumefaciens* linear chromosome (with bootstrap values of 98 and 78% for RepA1 and RepB1, respectively). (The nearest-neighbor analysis for RepC1 was inconclusive because of low bootstrap values.) The *min* genes play a role in the proper placement of the division septum during cell division (de Boer et al. 1989). We found that the pSymB *minCD* genes also share a common origin with the *minCD* genes found on the linear chromosome (100% bootstrap values), which are also likely to have been present in the ancestral plasmid because of their important function in cell division.

Synteny between the *A. tumefaciens* circular chromosome and the *S. meliloti* chromosome is preserved, with the exception of a large region containing approximately 300 genes; the *A. tumefaciens* orthologs of these genes are found on the linear chromosome and synteny is retained in small groups of genes (Goodner et al. 2001). Goodner et al. (2001) suggest that portions of the linear chromosome may have originated from an excision from the ancestral main chromosome, with subsequent insertion events and linearization. Given that pSymB and the *A. tumefaciens* linear chromosome have a common origin, it is likely that the excision occurred after the divergence of *S. meliloti* and *A. tumefaciens*, since this region is still intact in the main chromosome of *S. meliloti*. COG analysis showed that 36% of genes on the linear chromosome are orthologous to *S. meliloti* chromosomal genes, 25% are orthologous to pSymB genes, while only 12% are orthologous to pSymA genes (Wood et al. 2001). The remaining 27% were not orthologous to any *S. meliloti* genes, which suggests that significant gene loss and acquisition played a key role in the differentiation of these replicons. Since the divergence of these species, while the *S. meliloti* chromosome has remained fairly stable, the pSymB replicon has been substantially rearranged, and horizontal acquisition of genes has led to expansion of this replicon. The reason for

the instability of pSymB, relative to the main chromosome of *S. meliloti*, is unclear. It cannot be explained by the presence of IS or RIME elements, or A, B, C palindromic elements, since these make up a smaller proportion of pSymB than the chromosome. It is evident that horizontal transfer, rearrangements, and duplications of pSymB genes have played a major role in the evolution and adaptation of *S. meliloti*.

## Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (Canada) Research, Strategic, and Genomics grants to GBG.

## References

- Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L., and Ramuz, M. 1993. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J. Bacteriol.* **175**: 7869–7874.
- Burge, C., Campbell, A. M., and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* **89**: 1358–1362.
- Chain, P. S., Hernández-Lucas, I., Golding, B., and Finan, T. M. 2000. *oriT*-directed cloning of defined large regions from bacterial genomes: identification of the *Sinorhizobium meliloti* pExo megaplasmid replicator region. *J. Bacteriol.* **182**: 5486–5494.
- de Boer, P. A., Crossley, R. E., and Rothfield, L. I. 1989. A division inhibitor and a topological specificity factor coded for by the *minicell* locus determine proper placement of the division septum in *E. coli*. *Cell* **56**: 641–649.
- Deng, W., Gordon, M. P., and Nester, E. W. 1995. Sequence and distribution of *IS1312*: evidence for horizontal DNA transfer from *Rhizobium meliloti* to *Agrobacterium tumefaciens*. *J. Bacteriol.* **177**: 2554–2559.
- Eisen, J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**: 163–167.
- Ermolaeva, M. D., White, O., and Salzberg, S. L. 2001. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29**: 1216–1221.
- Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package), version 3.5c. University of Washington, Seattle, Washington.
- Finan, T. M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorhölter, F. J., Hernandez-Lucas, I., Becker, A., Cowie, A., Gouzy, J., et al. 2001. The complete sequence of the 1,683-kb pSymB megaplasmid from the  $N_2$ -fixing endosymbiont *Sinorhizobium meliloti*. *Proc. Natl. Acad. Sci. U.S.A.* **98**: 9889–9894.
- Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., et al. 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**: 668–672.
- García-Vallve, S., Romeu, A., and Palau, J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10**: 1719–1725.
- Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725–736.

- Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., et al. 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**: 2323–2328.
- Huynen, M., Snel, B., Lathe III, W., and Bork, P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**: 1204–1210.
- Hynes, M., Simon, R., Müller, P., Niehaus, K., Labes, M., and Pühler, A. 1986. The two megaplasmids of *Rhizobium meliloti* are involved in the effective nodulation of alfalfa. *Mol. Gen. Genet.* **202**: 356–362.
- Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16**: 332–346.
- Jukes, T. H. and Kimura, M. 1984. Evolutionary constraints and the neutral theory. *J. Mol. Evol.* **21**: 90–92.
- Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., Ramuz, M., and Allardet-Servent, A. 1998. Unconventional genomic organization in the alpha subgroup of the Proteobacteria. *J. Bacteriol.* **180**: 2749–2755.
- Karlin, S. and Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**: 283–290.
- Karlin, S., Ladunga, I., and Blaisdell, B. E. 1994. Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. U.S.A.* **91**: 12837–12841.
- Karlin, S., Mrázek, J., and Campbell, A. M. 1997. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**: 3899–3913.
- Koski, L. B. and Golding, G. B. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**: 540–542.
- Koski, L. B., Morton, R. A., and Golding, G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* **18**: 404–412.
- Lawrence, J. G. 1995. Ameliorator, Version 1.0.
- Lawrence, J. G. and Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**: 383–397.
- Lawrence, J. G. and Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U.S.A.* **95**: 9413–9417.
- Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., and Danchin, A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**: 851–856.
- Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**: 23–36.
- Muto, A. and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A.* **84**: 166–169.
- Østeras, M., Boncompagni, E., Vincent, N., Poggi, M. C., and Le Rudulier, D. 1998. Presence of a gene encoding choline sulfatase in *Sinorhizobium meliloti* bet operon: choline-O-sulfate is metabolized into glycine betaine. *Proc. Natl. Acad. Sci. U.S.A.* **95**: 11394–11399.
- Østeras, M., Stanley, J., and Finan, T. M. 1995. Identification of *Rhizobium*-specific intergenic mosaic elements within an essential two-component regulatory system of *Rhizobium* species. *J. Bacteriol.* **177**: 5485–5494.
- Palmer, K. M., Turner, S. L., and Young, J. P. 2000. Sequence diversity of the plasmid replication gene *repC* in the *Rhizobiaceae*. *Plasmid* **44**: 209–219.
- Pesole, G., Bozzetti, M. P., Lanave, C., Preparata, G., and Saccone, C. 1991. Glutamine synthetase gene evolution: a good molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **88**: 522–526.
- Sicheritz-Pontén, T. and Andersson, S. G. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**: 545–552.
- Strimmer, K. and von Haeseler, A. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**: 964–969.
- Suominen, L., Roos, C., Lortet, G., Paulin, L., and Lindstrom, K. 2001. Identification and structure of the *Rhizobium galegae* common nodulation genes: evidence for horizontal gene transfer. *Mol. Biol. Evol.* **18**: 907–916.
- Tabata, S., Hooykaas, P. J., and Oka, A. 1989. Sequence determination and characterization of the replicator region in the tumor-inducing plasmid pTiB6S3. *J. Bacteriol.* **171**: 1665–1672.
- Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.* **9**: 678–687.
- Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**: 33–36.
- Thomas, C. M. 2000. Paradigms of plasmid organization. *Mol. Microbiol.* **37**: 485–491.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- Turner, S. L. and Young, J. P. 2000. The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. *Mol. Biol. Evol.* **17**: 309–319.
- Wang, B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. *J. Mol. Evol.* **53**: 244–250.
- Wong, K., Finan, T. M., and Golding, G. B. 2002. Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids. *Funct. Integr. Genomics*, in press.
- Wood, D. W., Setubal, J. C., Kaul, R., Monks, D. E., Kitajima, J. P., Okura, V. K., Zhou, Y., Chen, L., Wood, G. E., Almeida NF, J. r., et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317–2323.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.
- Young, J. P. W. and Johnston, A. W. B. 1989. The evolution of specificity in the legume-rhizobium symbiosis. *TREE* **4**: 341–349.