J Mol Evol (2000) 51:459–463
DOI: 10.1007/s002390010108

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 2000

Replication Orientation Affects the Rate and Direction of Bacterial Gene Evolution

Elisabeth R.M. Tillier, Richard A. Collins

Department of Molecular and Medical Genetics, University of Toronto, 1 King's College Circle, Toronto, Canada, M5S 1A8

Received: 27 April 2000 / Accepted: 1 August 2000

Abstract. In many bacterial genomes, the leading and lagging strands have different skews in base composition; for example, an excess of guanosine compared to cytosine on the leading strand. We find that *Chlamydia* genes that have switched their orientation relative to the direction of replication, for example by inversion, acquire the skew of their new "host" strand. In contrast to most evolutionary processes, which have unpredictable effects on the sequence of a gene, replication-related skews reflect a directional evolutionary force that causes predictable changes in the base composition of switched genes, resulting in increased DNA and amino acid sequence divergence.

Key words: Base composition — Skews — Bacterial genomes — Replication orientation — Evolutionary force — Evolutionary rate.

Introduction

Many forces act directly or indirectly on a gene and ultimately determine its sequence. Of primary importance is selection for function of the gene product in its particular environment, which includes not only the exterior environment of the organism but also the internal environment of the cell and the genome itself. At the molecular level, DNA sequences are subject to the constraints of the interactions with other molecules, the genetic code, codon usage preferences, and base composition preferences (i.e., G+C content). Sequences are also subject to the stochastic effects of mutation. Recently, analyses of complete sequences of several bacterial genomes have revealed another constraint: the asymmetries in base composition with respect to the replication orientation (Lobry 1996a; McLean et al. 1998; Mrázek and Karlin 1998; Rocha et al. 1999; Tillier and Collins 2000). In most cases, the leading strand in replication is rich in guanosine (G) and thymidine (T) and the lagging strand is rich in cytosine (C) and adenosine (A), resulting in sequences having GC and AT skews. Replication orientation significantly affects gene sequences, independent of any coding biases, such as from amino acid or codon preferences (Rocha et al. 1999; Lafay et al. 1999; Tillier and Collins 2000; Romero et al. 2000).

To investigate the effect of base composition skews on the evolution of genes, we performed pairwise comparisons of genomes and analyzed the sequences of those genes that have switched orientation with respect to replication direction. For this, it was necessary to compare genomes that are similar enough that many orthologous gene pairs can be clearly identified, but sufficiently diverged that many genomic rearrangements and base substitutions have occurred. If the genomes have similar G+C content and GC skew, then differences in base composition of switched genes can be attributed solely to the difference in replication direction. The sequenced *Chlamydia* genomes presented many examples of orthologs that had switched replication orientation and the consequences for the evolution of such genes were then analyzed.

Correspondence to: E.R.M. Tillier; e-mail e.tillier@utoronto.ca

Materials and Methods

The complete genome sequences of *Mycoplasma genitalium, M. pneumoniae, Helicobacter pylori* 26695 and J99, *C. trachomatis* Serovar D, and *C. pneumonia* CWL029 and their predicted coding regions were obtained from the NCBI FTP server *ftp.//ncbi.nlm.nih.gov/genbank/genomes/bacteria/.*

FASTA 2 with BLOSUM50 matrix and with gap penalties of −12, −2 was used to find homologs in pairwise comparisons of the predicted coding regions of genomes (Pearson 1990). Two coding regions were considered orthologous and unique if they shared more than 30% amino acid identity with a probability of less than 0.01 and for which no other gene was found meeting these criteria and having at least half the z-score of the better match. Uniqueness is an important criteria to eliminate repeated sequences and paralogs.

The GC and AT skews are defined by the quantities (G−C)/(G+C) and (A−T)/(A+T) respectively (Blattner et al. 1997), and have been used to measure the preference for G over C and A over T, of the strand replicated continuously (the leading strand), or of its complimentary lagging strand in bidirectional replication of the chromosome. To determine the replication orientation of genes we inferred the location of the origin and termination of replication in the two genomes from the position of change in direction of the GC skew at third positions of codons (Lobry 1996b, 1996c). Genes within 1% (in terms of the total sequence length) of these were not considered because the origins and termination points have not been experimentally defined and these points are probably not exact.

When appropriate, two-sample *t*-tests (not assuming equal variances) were performed to evaluate the statistical significance of the differences in base composition and amino acid divergence of switched versus nonswitched genes. In some cases, the distributions tested are not normal and do not have the same shape. Since lack of normality violates the assumptions of the *t*-test and the different shapes (and thus variance) that of the Mann-Whitney test, we also used another nonparametric dominance test (Cliff 1993) to determine if the differences in skews for switched genes were significantly greater than that for nonswitched genes with 95% confidence. The dominance statistic is given by d = Probability (P1 > P2) − Probability (P2 > P1), where Probability (P1 > P2) is the number of times a score from population 1 is greater than a score from population 2 divided by the product of the populations' sizes.

Results and Discussion

Of the closely related sequenced genomes available, we found no switched genes in the two *Mycoplasma* species (*M. genitalium* and *M. pneumoniae;* Fraser et al. 1995; Himmelreich et al. 1996). The two *H. pylori* (26695 and J99) (Alm et al. 1999; Tomb et al. 1997) are too similar in sequence identity and have few clearly identifiable "switched" genes (we identified only 10 from 1159 orthologous pairs). Lafay et al. (1999) analyzed switched genes the two spirochete genomes *Borrelia burgoderfori* and *Treponema pallidum,* but these genomes are highly diverged and have very different base composition. The two sequenced *Chlamydia* genomes (*C. trachomatis* and *C. pneumoniae*) proved ideal for this analysis: these genomes share many homologous genes with a high level of synteny between the two genomes (Stephens et al. 1998; Kalman et al. 1999). Several genomic rearrangements are present, leading to many "switched" genes. Both genomes have the same G+C content (41%) but have large GC skews (0.13 and 0.09, respectively on the leading strand; Tillier and Collins 2000).

We identified 732 unique orthologous pairs of genes using FASTA analysis of the predicted coding regions of *C. pneumoniae* versus those of *C. trachomatis.* For 47 of these orthologs, the coding strand (mRNA synonymous strand) was on the leading strand in one genome and on the lagging strand in the other; these genes were defined as "switched" (Table 1). Several observations indicate that switched genes are functional: (1) They were identified by the same criteria as nonswitched genes (Stephens et al. 1998; Kalman et al. 1999); (2) unique orthologs of were found in the both genomes and include the only copy of known genes, such as those encoding pyruvate kinase, triosephosphate isomerase, and DNA topoisomerase I; (3) orthologs are approximately the same length in both genomes, indicating that, despite substantial sequence divergence, premature stop codons have been selected against. For simplicity, we will refer to the other 685 genes that are located on the same strand in both genomes as "nonswitched." For some analyses the nonswitched genes were divided into two groups: 281 genes that had moved to a different location but still on the same strand with respect to replication direction, defined as "moved"; and 404 genes that were located at the same relative position in both genomes, defined as "nonmoved." Because the moved and switched genes have both been subject to some kind of genome rearrangement process, separate analysis of the moved genes provides a way to determine if changes in base composition might have resulted simply from processes of rearrangement.

Figure 1 shows the distributions of the GC skew at third codon positions for the switched and nonswitched *Chlamydia* genes. Third codon positions were chosen for this figure because substitutions at these positions are the least likely to be influenced by additional effects of selection for amino acid preference due to the redundant nature of the genetic code and show the largest replication-related skews (Tillier and Collins 2000). On average, nonswitched genes on the leading strand in both genomes show a large positive GC skew (solid symbols); those on the lagging strand have a large negative skew (open symbols). Switched genes (dashed lines) have, on average, the skew of their current host strand. Although this analysis cannot determine in which genome the gene has actually changed orientation, nor the extent of change in any given gene, Fig. 1 suggests that the switched genes have undergone a change in base composition to acquire the skew of the new host strand.

To examine the extent to which base composition has changed as a result of switching strands, we performed pairwise comparisons of GC skews. Figure 2 shows the distributions for the difference between the GC skews (at each of the three codon positions) of orthologs in the two

Table 1. Switched genes in *C. trachomatis* and *C. pneumoniae*

C. pneumoniae PID	Length (aa)	*C. trachomatis* PID	Length (aa)	Gene name	Identity (%)
4376365	401	**3328406**	434	CT017 hypothetical protein	58.4
4376366	274	**3328405**	243	CT016 hypothetical protein	60.1
4376402	252	**3328505**	252	ACR family	66.7
4376404	189	**3328619**	190	YqgE hypothetical protein	58.2
4376405	147	**3328620**	149	YqdE hypothetical protein	59.2
4376406	232	**3328621**	243	Ribose-5-P isomerase A	53.4
4376490	229	**3328540**	229	YpdP hypothetical protein	49.3
4376492	373	**3328599**	373	Queuine tRNA ribosyl transferase	75.3
4376535	297	**3328629**	299	YqfU hypothetical protein	79.8
4376544	244	**3328536**	240	Lysophospholipase esterase	52.1
4376590	123	**3328497**	120	Acyl-carrier protein synthase	51.7
4376688	160	**3328499**	154	CT102 hypothetical protein	44.8
4376724	317	**3328395**	317	CT007 hypothetical protein	58.0
4376769	286	**3328807**	279	Deoxyheptonate aldolase	60.6
4376770	65	**3522894**	64	CT382.1 hypothetical protein	43.8
4376899	328	**3328921**	315	Ferrochetalase	53.7
4376989	218	**3328918**	218	CT482 hypothetical protein	47.2
4376990	253	**3328917**	244	CT481 hypothetical protein	48.0
4377021	100	**3329106**	98	CT656 hypothetical protein	49.0
4377022	107	**3329107**	106	CT657 hypothetical protein	38.7
4377023	326	**3329108**	335	Predicted pseudouridine synthase	64.4
4377024	79	**3329109**	79	CT659 hypothetical protein	86.1
4377070	354	**3329099**	353	RecA recombination protein	83.0
4377071	181	**3329098**	179	Formyltetrahydrofolate cycloligase	36.3
4377072	429	**3329097**	425	CT648 hypothetical protein	61.4
4377073	195	**3329096**	193	CT647 hypothetical protein	30.6
4377074	464	**3329095**	460	CT646 hypothetical protein	32.6
4377077	872	**3329092**	858	DNA Topoisomerase I	70.9
4377101	981	**3329033**	955	CT590 hypothetical protein	58.2
4377183	193	**3329182**	185	Biotin synthetase	53.5
4377376	383	**3329241**	378	Oxononanoate synthase	33.6
4376358	485	3328751	486	Pyruvate kinase	72.4
4376368	429	3328404	435	ATPase	74.1
4376536	193	3328628	193	Phenylacrylate decarboxylase	64.2
4376537	298	3328627	303	Benzoate octaphenyltransferase	41.3
4376666	167	3328435	168	CT043 hypothetical protein	94.0
4376765	219	3328808	244	CT383 hypothetical protein	34.2
4376777	340	3328815	409	CT389 hypothetical protein	68.5
4376975	365	3328993	326	RNA methyltransferase	54.3
4377075	99	3329094	99	CT645 hypothetical protein	64.6
4377076	336	3329093	335	Predicted oxidoreductase	69.3
4377078	265	3329090	272	CT642 hypothetical protein	55.8
4377148	155	3328996	160	CT556 hypothetical protein	68.4
4377184	416	3329183	380	Rod shape protein	74.5
4377216	77	3329214	75	CT753 hypothetical protein	53.3
4377279	451	3329272	451	CT805 hypothetical protein	64.1
4377399	255	3328746	275	Triosephosphate isomerase	47.1

Protein identification numbers and gene names are given for the 47 genes in *C. pneumoniae* and *C. trachomatis* with one ortholog on the leading strand (in bold) and the other ortholog on the lagging strand. The amino acid identity between the two orthologs is also given

genomes. A small difference in the GC skew was observed at all codon positions even for orthologs on the same strand in both genomes, as expected because *C. trachomatis* genes have, on average, larger GC skews than *C. pneumoniae* (Fig. 1 legend). For example, if a typical gene on the leading strand in *C. trachomatis* (average third codon position skew of +0.25) remained on the leading strand in *C. pneumoniae* (skew of +0.10) its skew would change by only 0.15. In contrast, if it switched to the lagging strand of *C. pneumoniae* (skew of −0.17) its skew would change by 0.42. We found that the distributions of skews at all codon positions of genes that have switched is statistically significantly different from those that have not switched (Fig. 2, see Materials and Methods); at third codon positions the differences are so large on average that the shape of the distribution is completely different for switched and nonswitched genes (Fig. 2C). The direction of the skew changes at all codon positions is in the expected direction, i.e., an increase in G on the leading strand and in C on the

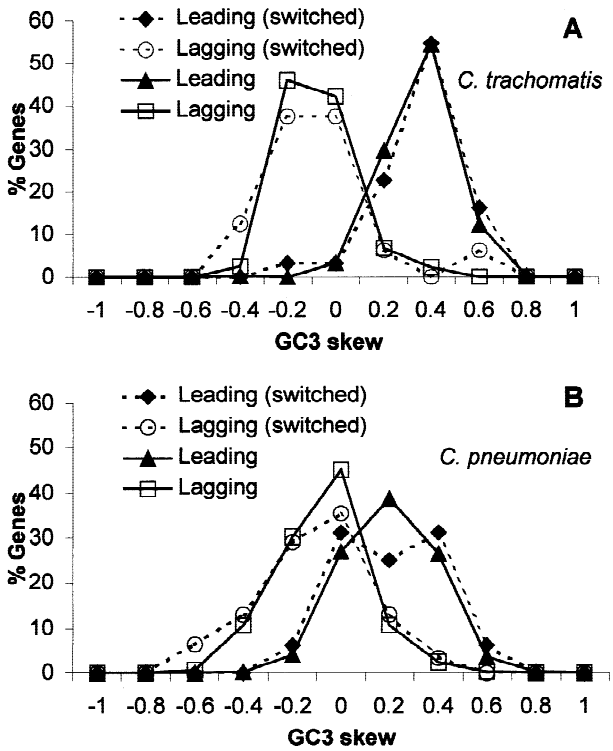

Fig. 1. Switched genes have the skew of their host strand. The distributions of the GC skew measured at the third position of codons (GC3) for nonswitched genes on the leading (solid symbols) and lagging strands (open symbols). The distributions for switched genes are given by corresponding solid and dashed lines. The average third position GC skew (± standard deviation) for nonswitched and switched genes on the leading and lagging strands in *C. trachomatis* are: 0.25 (± 0.13), 0.26 (± 0.15), −0.18 (± 0.15), −0.17 (± 0.20); and in *C. pneumoniae* are: 0.10 (± 0.18), 0.11 (± 0.21), −0.17 (±0.18), −0.22 (± 0.22).

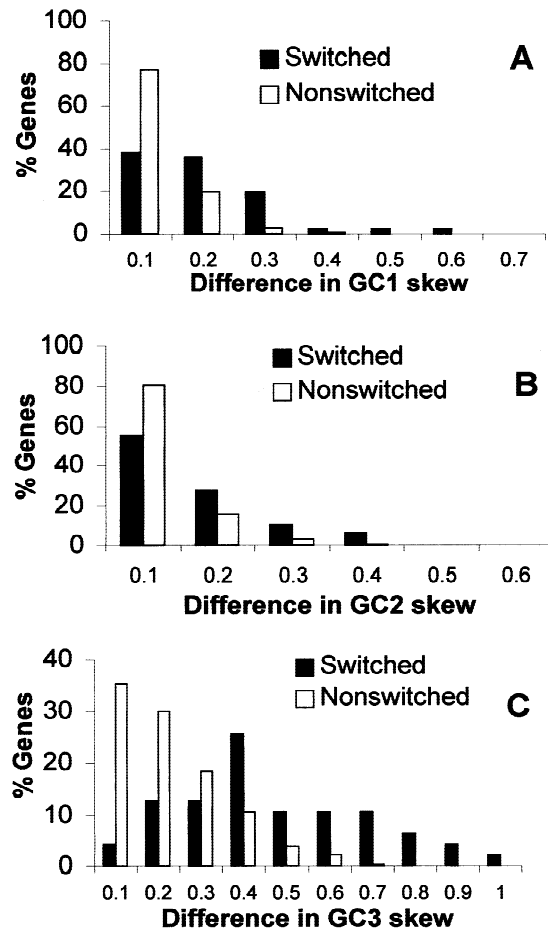

Fig. 2. Switched genes have a larger difference in GC skew than nonswitched genes. The distribution of the absolute value of the difference in the GC skew calculated at the first **(A),** second **(B),** and third **(C)** codon positions between the genes in *C. trachomatis* and their orthologs in *C. pneumoniae.* Filled and open bars show the distributions for switched and nonswitched orthologs, respectively.

lagging strand (as shown in Fig. 1 for the third positions, and data not shown for positions 1 and 2). In contrast, there is no significant difference between in the skews of genes that have moved elsewhere on the same strand compared to those that have not moved (data not shown), indicating that switching strands, not simply transposing to a new location, is the cause of the large change in skew observed for switched genes. The force to maintain base composition skews is thus strong enough that genes that have inverted their orientation with respect to the origin of replication appear to quickly accumulate mutations toward the opposite skew. This leads to greater base composition differences and therefore a greater decrease in average DNA sequence identity between genes where one ortholog has switched strands.

Because there is a significant difference in base composition skew at all codon positions between switched and nonswitched genes, we investigated whether these differences might be substantial enough to affect amino acid sequences between orthologs on different strands. We investigated both amino acid identity and amino acid similarity. Amino acid similarity counts the number of identical amino acids as well as amino acids that are similar in their properties and often interchanged in an alignment of two proteins sequences. The BLOSUM50 matrix used by FASTA (see Materials and Methods) defines the degree to which amino acids are similar.

Overall, the average amino acid identity for all the genes considered is 63.2% (standard deviation: ± 15%) and the average amino acid similarity (see Materials and Methods) is 83.5% (± 9%) and the distributions of identity and similarity are essentially identical for the moved and nonmoved subsets of the nonswitched genes (Fig. 3, open and gray bars, respectively, and data not shown). In contrast, switched genes are on average more diverged (black bars), with an average amino acid identity of only 57.8% (± 14.5%) and average similarity of 79.9% (± 10%) corresponding to statistically significant reductions of 8% and 4.3% in amino acid identity and similarity from genes that are in the same location or have moved to another position on the same strand.

The amino acid identity and similarity of moved genes do not differ significantly from those of nonmoved genes ($p = 0.12$ for identity and $p = 0.26$ for similarity,

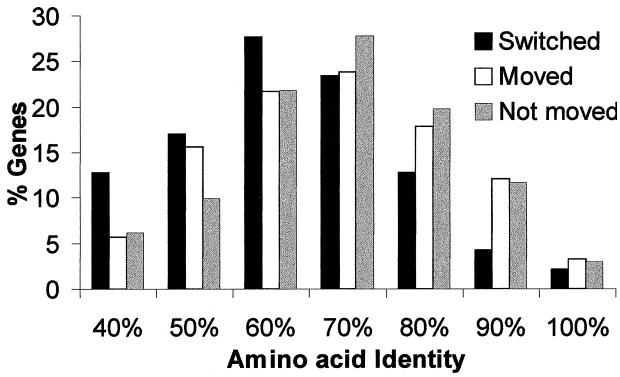

Fig. 3. Switching replication orientation increases amino acid sequence divergence. Distributions of the percentage amino acid identity between orthologs of *C. trachomatis* and *C. pneumoniae* that have switched orientation with respect to the direction of replication (black bars), between those orthologs on the same strand but that have moved in the genome (open bars), between those orthologs on the same strand but that have not moved in the genome (gray bars).

see Materials and Methods). However, comparisons of switched genes to nonswitched genes were statistically significant ($p = 0.005$ for the difference in amino acid identity and $p = 0.006$ for the difference in amino acid similarity). Comparison of switched genes to moved genes were also significant ($p = 0.017$ for the difference in amino acid identity, and $p = 0.010$ for the difference in amino acid similarity). This indicates that the amino acid identity and similarity for switched genes are significantly lower than those of nonswitched genes (moved or not).

Switching strands therefore has led to reductions in amino acid identity and similarity that are significantly greater than observed for nonswitched genes. An example of two homologous genes in *B. burgdorferi* where one paralog is on the leading strand and the other on the lagging strand has also suggested an influence of the replication direction on the amino acid sequence (Rocha et al. 1999). Lafay et al. (1999) also showed that orthologs in the two spirochetes *B. burgdorferi* and *T. pallidum* have the codon and amino acid usage of the host species appropriate for the replication direction. We find that the influence of the replication-related skew is strong enough to affect the average amino acid divergence over all 47 switched genes in the *Chlamydia* genomes. Changing the replication orientation of a gene therefore changes not only the base composition but also the amino acid sequence in a significant manner.

Mutational pressure leading to base composition differences between bacterial genomes have long been recognized (Sueoka 1962), but base compositional asymmetries between the leading and lagging strands has been a surprising discovery made possible by the complete sequencing of bacterial genomes. We show here that the force that maintains base composition skews is strong enough to affect the evolution of these sequences, even in the face of other selective forces, such as function, regulation, and codon and amino acid preferences. Most bacterial genomes show some base composition skews, and the degree to which the evolution of sequences is affected would be expected to be correlated to the magnitude of the replication-related skew of the host genomes. Switching strands not only leads to an increased evolutionary rate but the substitutions are in predictable direction (on average) toward a particular base composition.

Acknowledgments. R.A.C. is a member of the Canadian Institute for Advanced Research (CIAR). This work was funded by the National Sciences and Engineering Research Council of Canada (NSERC) grant to R.A.C.

References

Alm RA, Ling L-S, Moir DT, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* Nature 397:176–180

Blattner FR, Plunkett G, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Cliff N (1993) Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull 114:494–509

Fraser CM, Gocayne JD, White O, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Himmelreich R, Hilbert H, Plagens H, et al. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420–4449

Kalman S, Mitchell W, Marathe R, et al. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* Nature Genetics 21:385–389

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–1649

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strand of bacteria. Mol Biol Evol 13:660–665

Lobry JR (1996b) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie 78: 323–326

Lobry JR (1996c) Origin of replication of *Mycoplasma genitalium.* Science 272:745–746

McLean M, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation and gene orientation in 12 prokaryotic genomes. J Mol Evol 47:691–696

Mrázek J, Karlin S (1998) Strand composition asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA 95:3720–3725

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Rocha EPC, Danchin A, Viari A (1999) Universal replication biases in bacteria. Mol Microbiol 32:11–16

Romero H, Zavala A, Musto H (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucl Acids Res 28:2084–2090

Stephens RS, Kalman S, Lammel C, et al. (1998) Genome sequence of an obligate intracellular pathogen of humans, *Chlamydia trachomatis.* Science 282:754–759

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci 48:582–591

Tillier ERM, Collins RA (2000) The contributions of replication orientation, gene direction and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50:249–257

Tomb JF, White O, Kerlavage AR, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature 388: 539–547