circadian function for the *Neurospora* clock gene *frequency*. *Nature* 399, 584–586
19 Loros, J.J. and Feldman, J.F. (1986) Loss of temperature compensation of circadian period length in the *frq-9* mutant of *Neurospora crassa*. *J. Biol. Rhythms* 1, 187–198
20 Lakin-Thomas, P.L. (2000) Circadian rhythms: new functions for old clock genes? *Trends Genet.* 16, 135–142
21 Emery, P. *et al.* (2000) A unique circadian-rhythm photoreceptor. *Nature* 404, 456–457
22 Bruce, V. (1960) Environmental entrainment of circadian rhythms. *Cold Spring Harbor Symp. Quant. Biol.* 25, 29–48

M. Merrow*
T. Roenneberg
Institute for Medical Psychology,
Goethestrasse 31, 80336 Munich, Germany.
*e-mail: martha@imp.med.uni-muenchen.de

Genome Analysis

Gene conversions in genes encoding outer-membrane proteins in *H. pylori* and *C. pneumoniae*

I. King Jordan, Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin

***Helicobacter pylori* and *Chlamydia pneumoniae* are both pathogenic to humans. Their genomes have recently been completed, allowing detailed study of their evolution and organization. Here we describe an evolutionary analysis of the *H. pylori* and *C. pneumoniae* genes that encode their outer-membrane proteins. By comparing complete genome sequences of two *H. pylori* strains and two *C. pneumoniae* strains, we identify multiple independent conversions among these genes. Such recombination events might provide a selective advantage for these bacterial pathogens.**

H. pylori is a Gram-negative, human-specific gastric pathogen, which is a causative agent of chronic active gastritis as well as duodenal and gastric ulcers[1]. Chronic *H. pylori* infection can also have a role in the development of gastric carcinomas[2]. *Chlamydia pneumoniae* is another a human pathogen, which causes bronchitis and pneumonia[3]. In addition, *C. pneumoniae* infection has been associated with atherosclerosis[4]. The availability of complete genomic sequences of two *H. pylori* strains[5,6] and two *C. pneumoniae* strains[7,8] allows for detailed inferences concerning the genome organization and evolution of these medically important organisms to be made. We have employed these genomic sequence data in an evolutionary analysis of *H. pylori* and *C. pneumoniae* gene families that encode outer-membrane proteins.

Examination of the complete *H. pylori* genome sequences revealed the presence of the large Hop family of outer-membrane proteins[5,9]. All Hop-family members contain a conserved C-terminal domain. Members of the Hop family were initially characterized as porins with similar N-terminal amino acid sequences[10,11]. Subsequently, additional Hop-family members were found to be involved in adhesion to the gastric endothelium[12–14]. The two sequenced *C. pneumoniae* genomes also encode polymorphic families of outer-membrane proteins[8]. For example, the *C. pneumoniae* CWL029 genome encodes 21 members of the outer-membrane-protein family[7]. The biological role of this family is unknown, but the patterns of variation among the genes of the family indicate that molecular mechanisms exist to promote functional diversity of their encoded products.

Many of these outer-membrane proteins are probably important in pathogenesis and the presence of such proteins encoded by repetitive gene families indicates a possible role for the families in antigenic variation and host-defense evasion[15]. Several different mechanisms involving recombination among repeated genes can influence antigenic variation. Gene conversion is an intragenomic, nonreciprocal recombination event that results in identical (homogenized) gene sequences[16]. In bacterial pathogens, gene conversion is thought to be important in the generation of the repertoire of 'contingency genes' that mediate pathogen–host interactions[15]. In particular, there is evidence that antigenic variation in *Neisseria gonorrhoeae* pilus proteins is shaped by gene conversion between pilus genes[17]. In addition, recombination between *Mycoplasma genitalium* dispersed repetitive elements and the Mga operon probably generates antigenic variation in cellular adhesin proteins that are required for attachment of the organism to host epithelium[18]. Tomb *et al.*[5] hypothesized that similar recombination mechanisms could contribute to genetic, and subsequently antigenic, variation of the Hop gene family and its encoded products.

Although conversion has been invoked as an important mechanism of antigenic-variation maintenance, rigorously distinguishing this recombination mechanism from very recent intragenomic duplication is difficult. The complete genome sequences of two *H. pylori* and two *C. pneumoniae* strains provide the data necessary explicitly to test the hypothesis that conversion occurs between copies of gene family members that encode

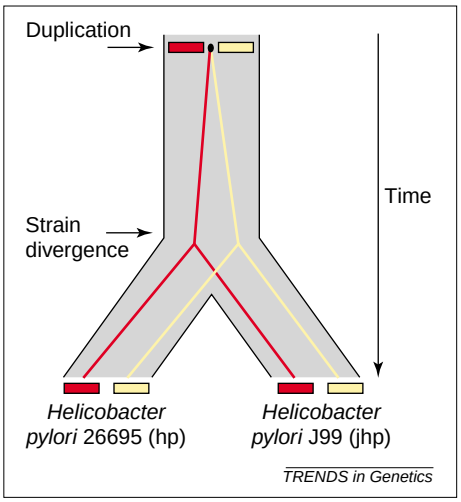

Fig. 1. Expected phylogenetic relationships among members of a gene family. Two strains, each containing two copies of a gene family, are represented. The strain lineage is shown in thick gray, and the gene lineages are shown with colored lines. Orthologs are indicated with the same color boxes and paralogs with different colors. Paralogous copies of a multi-gene family last shared a common ancestor at the time of gene duplication, whereas orthologous genes last shared a common ancestor at the time of strain divergence.

  PII: S0168-9525(00)02151-X

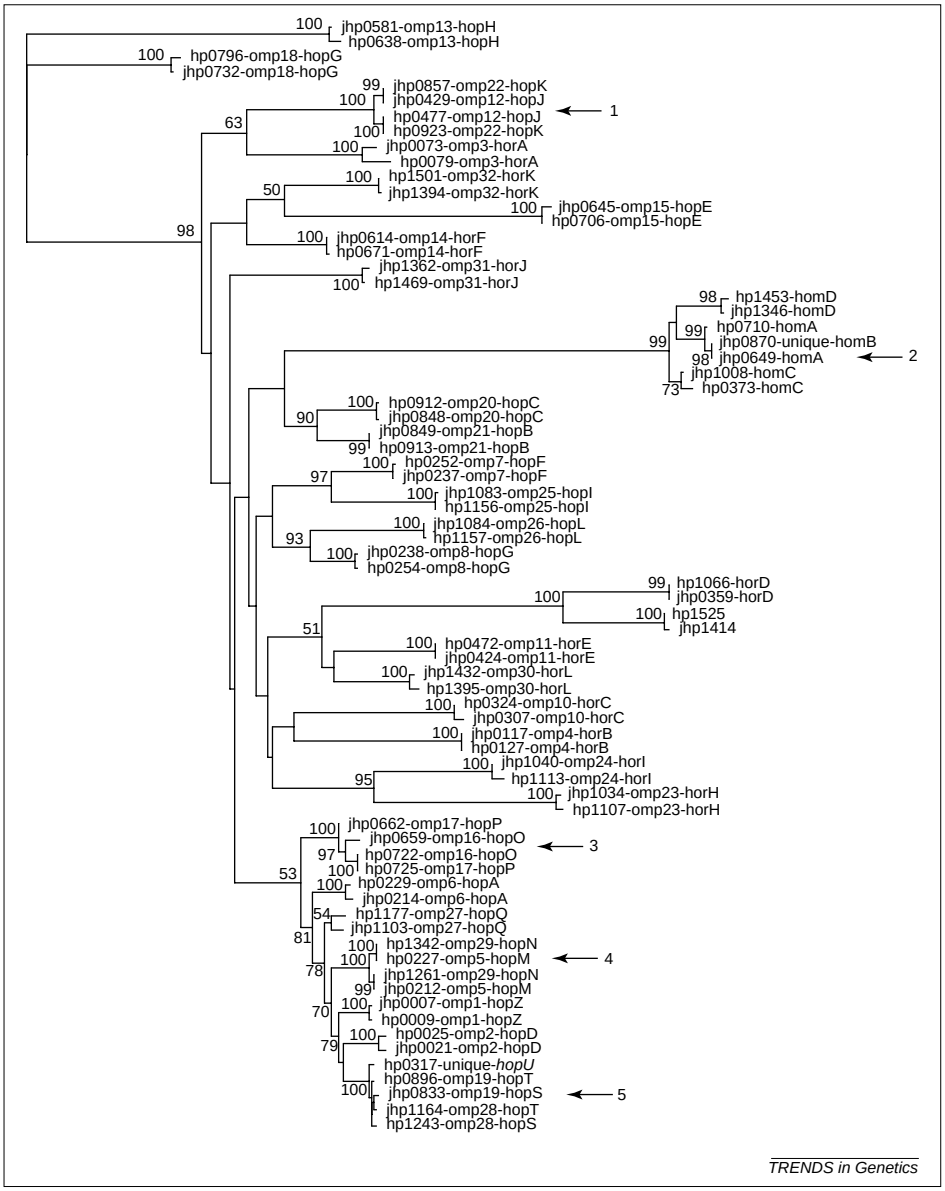

Fig. 2. Hop-family phylogeny. Taxa names consist of a *Helicobacter pylori* strain-specific locus number followed by the outer-membrane-protein (OMP) gene number and the gene name (as in Refs 5, 9). hp, *H. pylori* strain 26695; jhp, strain J99. Bootstrap values (100 replicates) are shown to the left of the node that they support. Five clades that show evidence of gene conversions are indicated with arrows. Hop protein sequences encoded by the two strains were identified using the BLAST[20] and COGnitor[21] programs.The SMART program[22] was used to identify the conserved C-terminal (OMP) domains of all Hop proteins. A multiple sequence alignment of the Hop–OMP domains generated by CLUSTALX[23] was used with the neighbor-joining[24] program of the PHYLIP package[25] in phylogenetic reconstruction of the family.

bacterial outer-membrane proteins. In particular, comparative analysis of intraspecific genomes allows for the crucial distinction between orthologous and paralogous members of gene families. Orthologs are genes in different species (or strains) that have evolved divergently from a common ancestor after speciation (or strain divergence), and paralogs are copies of a gene family that have evolved divergently from a common ancestor after gene duplication[19]. The approach employed here relies on the certainty that orthologs from two strains of the same species can be identified in complete genome sequences owing to the fact that they occupy the same chromosomal location. The presence of copies of a gene family located identically in the genomes of two strains indicates that the duplication events that generated the copies occurred before strain divergence (Fig. 1). Therefore, genes that are orthologous between strains are expected to be more closely related than genes that are paralogous within a genome, because the orthologs share a more recent common ancestor (at the time of strain divergence) than the paralogs (at the time of duplication).

The phylogenetic relationships (Fig. 2) among 36 orthologous pairs of Hop proteins and two 'unique' Hop proteins (encoded by genes that do not have orthologs in the corresponding strain) were analyzed to test this prediction. The majority of orthologous pairs (26) show phylogenetic relationships consistent with the expectations described above. However, there are five clades in the outer-membrane-protein tree where paralogous sequences are more closely related than orthologous sequences. Because the orthologous sequences share a more recent common ancestor than do the paralogs (Fig. 1), this evidence indicates that these paralogous sequences have been homogenized within genomes (strains). Each of these clades was analyzed further using both amino acid and nucleotide alignments of entire coding regions and flanking noncoding regions. The results reveal a complex pattern of nine apparent gene conversion events involving 18 Hop genes (eight in strain 26695 and ten in strain J99). Six of the *H. pylori* gene conversions (Fig. 3a–c) are paired in the sense that they involve sequence homogenization between the same pair of paralogs within the genomes of each strain. As these sequences differ between strains, these data indicate that conversion occurred independently in each strain, subsequent to the strain divergence. Four of these six paired events (Fig. 3a,b) are conversions of the entire coding region and some flanking sequence, and two cases (Fig. 3c) involve the generation of chimeric sequences that contain divergent 5′ regions and identical 3′ regions. In addition to the paired conversion events, there are three unpaired conversion events. Two of the unpaired events (Fig. 3e,f) involve homogenization between 'unique' Hop genes (hp0317 and jhp0870) and genes with corresponding orthologs in the other strain. Interestingly, both conversion events involving 'unique' genes are complex, with multiple conversion tracts.

Previously, it has been proposed that DNA taken up by natural transformation from surrounding lysed cells could result in intrastrain identity between paralogous Hop genes[9]. However, transformation is not a mechanism that

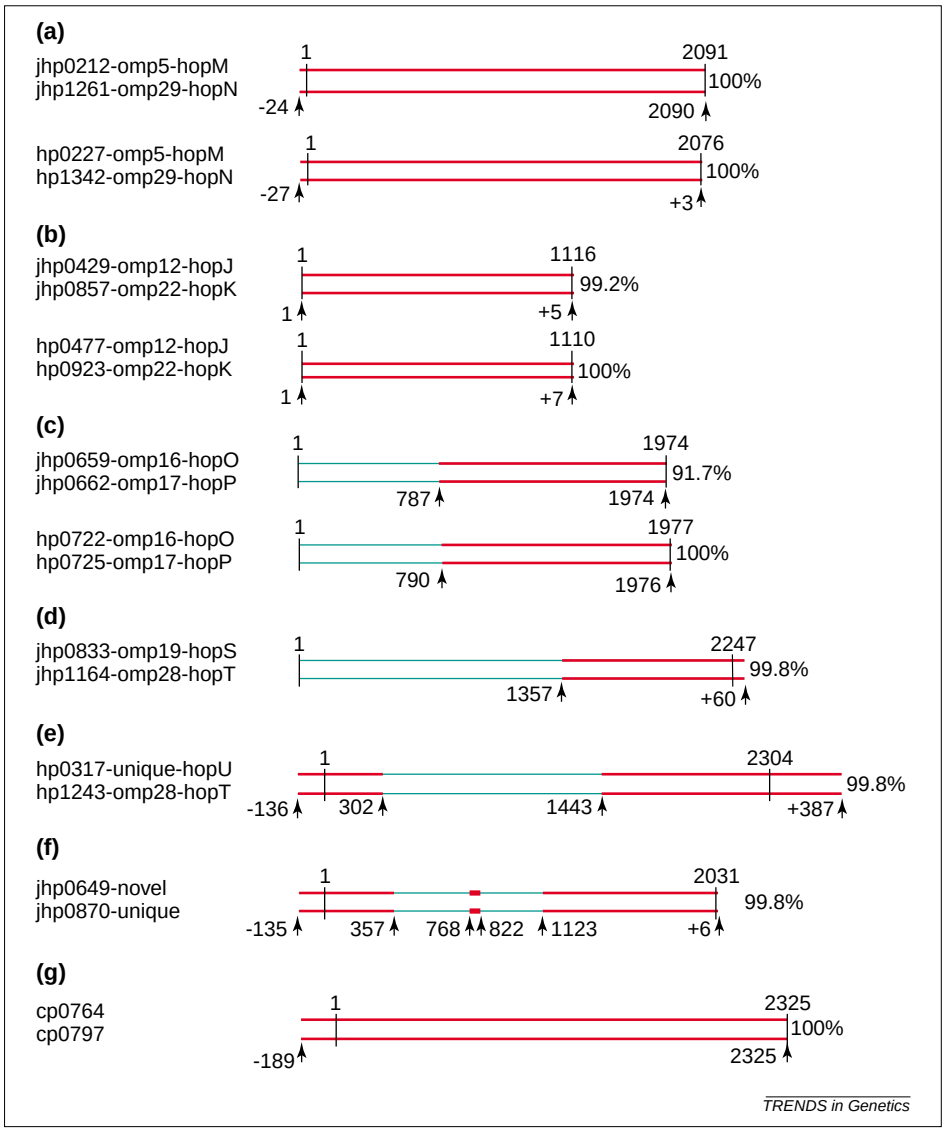

Fig. 3. Gene conversions. Pairs of genes involved in gene conversion are shown. Red, converted sequences; green, nonconverted sequences. Sequences are numbered with respect to the beginning and the end of the open reading frames (ORFs). Percent nucleotide identities within the conversion tracts are shown.

necessarily leads to intragenomic homogenization. For this mechanism to explain the data observed here, both paralogous genes would have to be transformed independently by the same sequence or one of the copies would have to be transformed by a sequence identical to the other copy. Furthermore, in the case of paired homogenization the same type of events (with different transforming sequences) would have to occur in both strains. Although transformation is a formal possibility, gene conversion seems a more probable mechanism, because it is known to homogenize sequences within genomes.

Paralogous gene families within the two *C. pneumoniae* genomes, including several that are predicted to encode outer-membrane proteins, were analyzed in the same way. Unlike the Hop family, which showed evidence of multiple gene conversions in both *H. pylori* strains, the *C. pneumoniae* genomes revealed evidence of only one gene conversion between paralogs (cp0764 and cp0797) encoding predicted outer-membrane proteins in strain AR39. This gene conversion (Fig. 3g) resulted in coding regions that show 100% nucleotide identity with one gap, and the gene conversion tract extends 189 bp upstream of the initiation codon. By contrast, the genes involved do differ between the two strains and the same two paralogs in *C. pneumoniae* strain CWL029 (cpn0010 and cpn1054) differ at 28 sites in the coding region and show no sequence identity outside of the open reading frames (ORFs); apparently, no gene conversion took place in this strain.

Intraspecific comparison of multiple complete genome sequences enabled a clear distinction between orthologs and paralogs, and thus facilitated explicit testing of a hypothesis concerning the mechanisms of generation of variation in outer-membrane proteins with potential antigenic significance. By simultaneously analyzing Hop gene families from both completely sequenced *H. pylori* strains, we obtained evidence of multiple independent gene conversions and of the generation of mosaic Hop genes within the genomes of both *H. pylori* strains. In addition, we identified one apparent case of gene conversion between two putative outer-membrane-protein encoding genes in *C. pneumoniae* strain AR39. The action of gene conversion on functionally similar (but unrelated) proteins in different species, and in particular the evidence of independent conversions between the same paralogs in different strains, strongly suggests that these recombination events provide a selective advantage for bacterial pathogens.

References

1 Dubois, A. (1995) Spiral bacteria in the human stomach: the gastric helicobacters. *Emerg. Infect. Dis.* 1, 79–85

2 McNamara, D. and O'Morain, C. (1998) *Helicobacter pylori* and gastric cancer. *Ital. J. Gastroenterol. Hepatol.* 30 (Suppl. 3) S294–298

3 Kuo, C.C., Jackson, L.A., Campbell, L.A. and Grayston, J.T. (1995) Chlamydia pneumoniae (TWAR). *Clin. Microbiol. Rev.* 8, 451–461

4 Kuo, C.C. *et al.* (1993) Demonstration of *Chlamydia pneumoniae* in atherosclerotic lesions of coronary arteries. *J. Infect. Dis.* 167, 841–849

5 Tomb, J.F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388, 539–547

6 Alm, R.A. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397, 176–180

7 Kalman, S. *et al.* (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* 21, 385–389

8 Read, T.D. *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28, 1397–1406

9 Alm, R.A. *et al.* (2000) Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infect. Immun.* 68, 4155–4168

10 Doig, P. *et al.* (1995) Isolation and characterization of a conserved porin protein from *Helicobacter pylori*. *J. Bacteriol.* 177, 5447–5452

11 Exner, M.M. *et al.* (1995) Isolation and characterization of a family of porin proteins from *Helicobacter pylori*. *Infect. Immun.* 63, 1567–1572

12 Odenbreit, S. *et al.* (1999) Genetic and functional characterization of the alpAB gene locus essential for the adhesion of *Helicobacter pylori* to human gastric tissue. *Mol. Microbiol.* 31, 1537–1548
13 Peck, B. *et al.* (1999) Conservation, localization and expression of HopZ, a protein involved in adhesion of *Helicobacter pylori*. *Nucleic Acids Res.* 27, 3325–3333
14 Ilver, D. *et al.* (1998) *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* 279, 373–377
15 Deitsch, K.W. *et al.* (1997) Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol. Mol. Biol. Rev.* 61, 281–293
16 Holliday, R. (1986) Gene conversion. *Prog. Clin. Biol. Res.* 218, 95–107
17 Haas, R. and Meyer, T.F. (1986) The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. *Cell* 44, 107–115
18 Peterson, S.N. *et al.* (1995) Characterization of repetitive DNA in the *Mycoplasma genitalium* genome: possible role in the generation of antigenic variation. *Proc. Natl. Acad. Sci. U. S. A.* 92, 11829–11833
19 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
20 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
21 Tatusov, R.L. *et al.* (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36
22 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234
23 Thompson, J.D. *et al.* (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882
24 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425
25 Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* 266, 418–427

I.K. Jordan*
National Center for Biotechnology Information, National Institutes of Health, Building 38A/Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA.
*e-mail: jordan@ncbi.nlm.nih.gov

K.S. Makarova
Uniformed Services University of the Health Sciences, Bethesda, MD 20894, USA.
Permanent address: Institute of Cytology and Genetics, Russian Academy of Sciences, Novisibirsk, 630090, Russia

Y.I. Wolf
E.V. Koonin
National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA.

Genome Analysis

Evolution of prokaryotic gene order: genome rearrangements in closely related species

Mikita Suyama and Peer Bork

Conservation of gene order in prokaryotes has become important in predicting protein function because, over the evolutionary timescale, genomes are shuffled so that local gene-order conservation reflects the functional constraints within the protein. Here, we compare closely related genomes to identify the rate with which gene order is disrupted and to infer the genes involved in the genome rearrangement.

Predicting protein function from the conservation of gene order is a method that complements more traditional homology-based methods (Refs 1–5 and references therein). Early measurements indicated that gene order is mostly disrupted if the average protein sequence identity of orthologs shared between two genomes is <50% (Ref. 1). Furthermore, gene order is randomized (except gene clusters with functional constraints) if the 16S rRNA distance measured by the number of substitutions per site exceeds 0.13 (Ref. 4). By comparing closely related genomes, we gained insights into the rate of disruption of gene order and which genes might be involved in the genome rearrangement.

Genome comparisons

We carried out 21 pairwise comparisons of genomes where the number of substitutions per site for 16S rRNA is <0.13 (see the legend of Fig. 1 for the genomes used). To study the evolution of gene order, orthologs in each genome pair had to be identified. We used the following conditions:

- candidates must have a homolog in the other genome identifiable by BLAST (Ref. 6) (using a cutoff expected rate of false positives of $E = 0.0001$);
- >80% of residues must be included in the BLAST alignment;
- both candidates must be the best hit to each other (reciprocal confirmation).

In this study we focused only on the orthologous genes between a pair of genomes.

Dotplots of the genome comparisons showed several patterns in genome rearrangement (Fig. 1). For example, we identified hot spots of genome rearrangement at the terminus of replication for ML, MT and VC1, in addition to those reported for EC, CP, CT, PH and PA (Refs 7–9; Fig. 1, see legend for abbreviations). Genome rearrangement at the terminus of replication is probably a general phenomenon in prokaryotes[10]. Furthermore, although the extent of inversions seems to vary in the species studied, most of them occur at the origin or terminus of replication (Fig. 1b,c,e–g,i). Thus, replication is linked not only to the rearrangement at the hot spots, but also to the inversion of large fragment of genomes. An extreme seems to be the lineage of proteobacteria exemplified by the EC versus VC1 comparison (Fig. 1g) where clusters of orthologs along the diagonals between the origin and terminus of replication indicate that multiple inversions pivoted on the origin or terminus of replication are the driving force of gene rearrangement. The other extreme is the absence of inversions in the mycoplasmas, despite their evolutionary distance (Figs 1h and 2; see also Box 1).

Neighborhood disruption frequencies

To quantify genetic processes and the patterns observed, we introduced a measurement, the neighborhood disruption frequency (NDF), that evaluates how gene order is conserved for a given genome pair. The NDF value is the number of measured breakpoints of gene neighbors[11] per number of shared genes between the genomes. The NDF ranges from 0 (complete conservation of gene order with no breakpoint) to 1 (complete shuffling). For example, the number of orthologous

 0168-9525/01/$ – see front matter PII: S0168-9525(00)02159-4