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 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 M. LT 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 prokaryotes10.

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). 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 genome 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 neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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 sequences is the number of measured breakpoints of gene neighbors11 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).
genes and the number of breakpoints of orthologous gene neighbors in the comparison of EC with VC1 are 1454 and 595, respectively, and thus the NDF value is 0.409.

We observed an almost linear increase of NDF against the number of substitutions per site (Fig. 2), with the exception of mycoplasmas and chlamydiads. Approximately 40% of the gene order of orthologs is disrupted at the evolutionary distance of 0.3 amino acid substitutions per site. The linear correlation indicates that not only the number of amino acid substitutions, but also the degree of genome rearrangement, constantly increases along the time of divergence. To identify genetic causes for the phenomenon that mycoplasmas and chlamydiads do not follow the general trends in genome evolution, we analyzed a number of possible reasons and distinct genetic features of these species (Box 1). Among the possible reasons, the lack of certain replication proteins seems the most plausible because the general trends of genome rearrangements are associated with replication (Fig. 1).

In summary, we show that there is a general tendency of rearrangement hot spots to be located near the terminus of replication and that most of the centers of inverted fragments are located at the terminus of replication. To a lesser extent, these tendencies are also true for chlamydia (CPa, CP, CT and CTm) and euryarchaeota (M). PA, PH and PF). These panels are ordered according to the number of amino acid substitutions per site for orthologous gene pairs (Fig. 2). The axes are graduated in 200 kb. Directional similarity is indicated by colors: green, pairs of genes with the same direction; red, those with opposite directions. The open reading frames (ORFs) without significant similarity to the other compared genome even in local DNA sequence level are defined as the species specific ORFs and indicated by blue dots on each axis. Species-specific ORFs are not identified for ML, because the ORFs of ML are determined by the orthology with MT. Arrows and lines indicate the predicted and/or experimentally determined origin (cyan) and terminus (pink) of replication. In the absence of experimental evidence for the terminus of replication, the site was predicted from the change in GC skew sign (data available on http://www.embl-heidelberg.de/~suyama/gene_order/index.html). Where there is no clear change in skew sign at the terminus, we predicted the terminus as the opposite site to the origin of replication.

Fig. 1. Dotplots of nine selected genome pairs. (a) Chlamydia pneumoniae AR39 (CPa) vs. C. pneumoniae CWLO29 (CP); (b) Neisseria meningitidis serogroup A strain Z2491 (NMa) vs. N. meningitidis serogroup B strain MC58 (NMb); (c) Helicobacter pylori 99 (HP99) vs. H. pylori 26695 (HP); (d) Chlamydia trachomatis serovar D (CT) vs. C. trachomatis MoPn (CTm); (e) Mycobacterium leprae (ML) vs. Mycobacterium tuberculosis (MT); (f) Pyrococcus horikoshii (PH) vs. Pyrococcus abyssi (PA); (g) Escherichia coli (EC) vs. Vibrio cholerae chromosome I (VC1); (h) Mycoplasma pneumoniae (MP) vs. Mycoplasma genitalium (MG); (i) CP vs. CT. Other genomes used in this study are Campylobacter jejuni (CJ), Haemophilus influenzae (HI), Methanococcus janaschii (MJ), and Pyrococcus furiosus (PF). All the genome sequences and data were obtained from GenBank, except for those for ML (The Sanger Centre; ftp://ftp.sanger.ac.uk/pub/pathogens/leprae) and PF (Utah Genome Center; http://www.genome.utah.edu/sequence.html), which are obtained through the web. These genomes correspond to different prokaryotic lineages: Gram-positive (MG, MP, ML and MT), proteobacteria (CJ, EC, HI, VC1, HP, HP99, NMa and NMb), respectively, and thus the NDF value is 0.409.

We observed an almost linear increase of NDF against the number of substitutions per site (Fig. 2), with the exception of mycoplasmas and chlamydiads. Approximately 40% of the gene order of orthologs is disrupted at the evolutionary distance of 0.3 amino acid substitutions per site. The linear correlation indicates that not only the number of amino acid substitutions, but also the degree of genome rearrangement, constantly increases along the time of divergence. To identify genetic causes for the phenomenon that mycoplasmas and chlamydiads do not follow the general trends in genome evolution, we analyzed a number of possible reasons and distinct genetic features of these species (Box 1). Among the possible reasons, the lack of certain replication proteins seems the most plausible because the general trends of genome rearrangements are associated with replication (Fig. 1).

In summary, we show that there is a general tendency of rearrangement hot spots to be located near the terminus of replication and that most of the centers of inverted fragments are located at the terminus of replication. To a lesser extent, these tendencies are also true for
the origin of replication. We also found an almost linear relationship between divergence of sequence and gene order degradation in closely related prokaryotic genomes. Mycoplasmas and chlamydias do not follow this linearity, and they both lack some of the genes involved in restarting the replication forks. We propose that these missing genes are the major factor for the slower rate of genome rearrangement in these organisms.

References

g Liu, J. and Marians, K.J. (1999) PriA-directed assembly of a primosome on D loop DNA. J. Biol. Chem. 274, 25033–25041
Correlation coefficient values calculated for the data points with and without dotted line, are shown in red. The outliers are indicated in blue. The rest of the genomes). See legend of Fig. 1 for abbreviations of the species.

Fig. 2. Distance-dependent genome shuffling. The horizontal axis denotes the averaged number of amino-acid substitutions/site for the orthologous genes shared among the 20 genomes analyzed in this study. The vertical axis indicates the neighborhood disruption frequency (NDF; number of orthologous gene neighborhood breakpoints per number of orthologs between a pair of genomes). See legend of Fig. 1 for abbreviations of the species.

Note added in proof
After submission of this paper, Tiller and Collins showed the impact on genome rearrangements of replication-directed translocation in some closely related organisms.

Acknowledgements
Wethank R. Copley, M. A. Huynen, W. C. Lathe III and B. Snel for critical reading of the manuscript.

References

Meeting Report
Embryo jigsaws
Stephen Kerridge

The Jacques Monod Conference on the Molecular and Cellular Basis of Morphogenesis was held in Aussois, France, from 7–11 October 2000.

Under the auspices of the Centre National de la Recherche Scientifique (CNRS) the Jacques Monod Conference was held surrounded by the snow-capped mountains of the French Alps. Superbly orchestrated by Dado Boncini and Michel Labouesse, scientists from all over the world met to unravel the mysteries of morphogenesis in animal development.

Cell polarity
Cells often organize into epithelia that possess inherent polarity with an apical, lateral and basal surface. Epithelia are held in sheets by their lateral surfaces which include different types of junctions that separate apical from lateral domains. What are the factors determining cell polarity? In

Caenorhabditis elegans, CHE14 encoding a twenty-pass transmembrane protein on the apical surface is required for epithelialization (Michel Labouesse, IGBMC, Strasbourg, France). The protein is required for ecyotaxis and shares similarity with the Drosophila Disched protein (better known for its role in Hedgehog secretion), which possesses a lipid-sensing domain perhaps required for targeting to the membrane. The worm surprisingly lacks

http://tig.trends.com 0168-9525/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02164-8