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

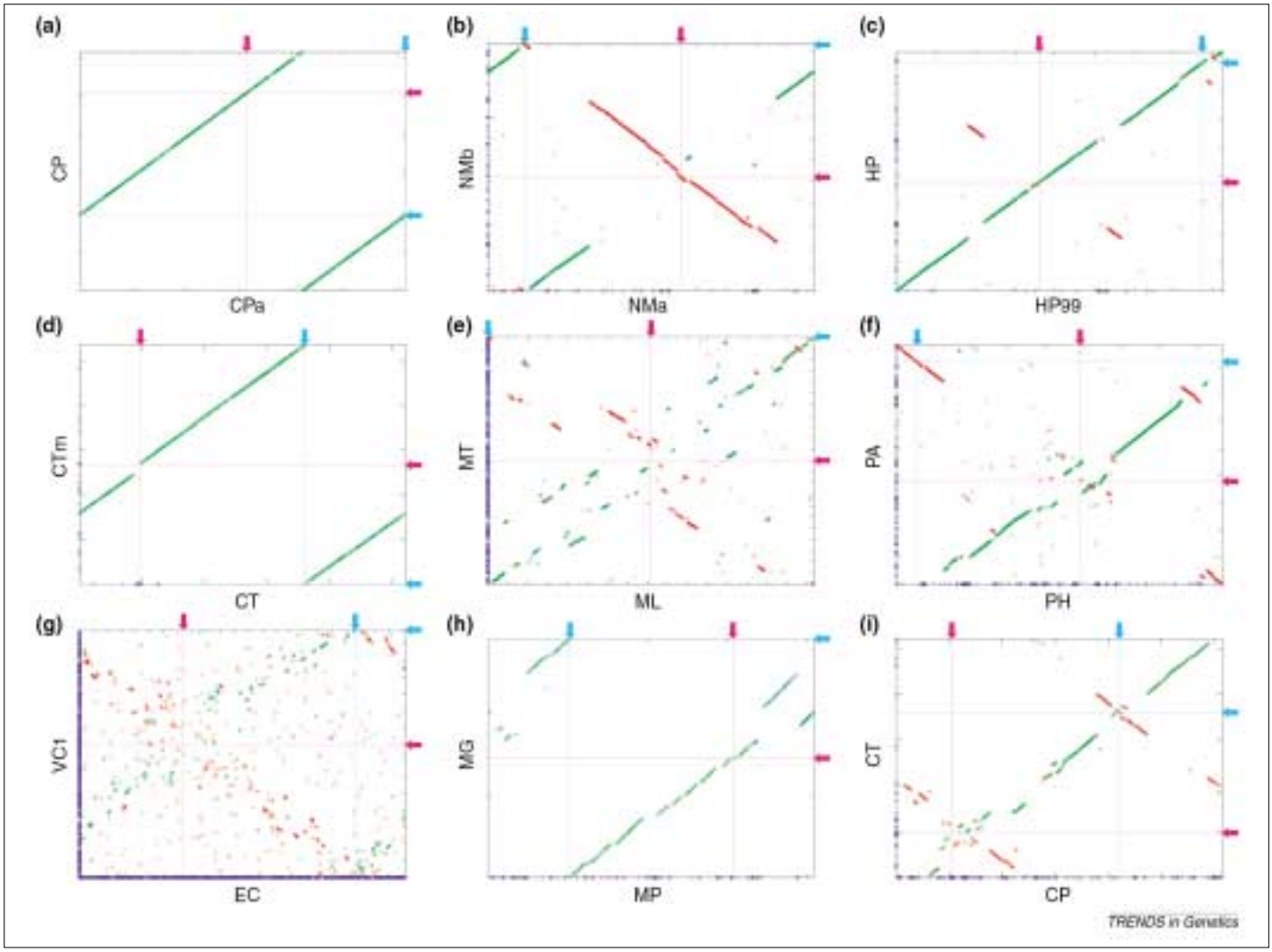

Fig. 1. Dotplots of nine selected genome pairs. (a) *Chlamydia pneumoniae* AR39 (CPa) vs. *C. pneumoniae* CWL029 (CP); (b) *Neisseria meningitidis* serogroup A strain Z2491 (NMa) vs. *N. meningitidis* serogroup B strain MC58 (NMb); (c) *Helicobacter pylori* J99 (HP99) vs. *H. pylori* 26695 (HP); (d) *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 1 (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 jannaschii* (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), chlamydia (CPa, CP, CT and CTm) and euryarchaeota (MJ, 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[8,9,12–17]. 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.

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 chlamydias. 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 chlamydias 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

Box 1. Possible reasons for the anomalous rate of genome rearrangement in mycoplasmas and chlamydias.

Restriction enzymes

Restriction enzymes might have a significant effect on genome rearrangement because they cut specific sites in DNA. Mycoplasmas and chlamydias analyzed in this study have no type I, type II or type III restriction enzymes, although MP has a frameshifted R-subunit of type I enzyme and also contains the rest of the subunits of the a type I enzyme. There is a relationship between avoidance of particular palindromic subsequences in a genome and the presence of certain restriction enzymes[a]. However, mycobacteria, which fit well with the linearity in Fig. 2, neither contain restriction enzymes nor show a significant avoidance of palindromic sequence (data available on http://www.embl-heidelberg.de/~suyama/gene_order/index.html). Thus, absence of restriction enzymes alone does not explain the low rearrangement rate for mycoplasmas and chlamydias.

Faster mutation rate

Some DNA-repair systems are not present in mycoplasmas and chlamydias[b]. Such a deficiency of repair systems might cause high mutation rates for these species, resulting in the points in Fig. 2 being further to the right than expected from the real divergence time. To check this effect we carried out phylogenetic analysis using the weighted neighbor-joining[c] method for 16S rRNA, EF-Tu/1α and EF-G/2 genes (data available on http://www.embl-heidelberg.de/~suyama/gene_order/index.html). Only the mycoplasmas show slightly higher mutation rates than other eubacteria; that is, this feature provides only a partial explanation.

Missing proteins required for genome rearrangement

Missing genes in particular genomes were identified using the cluster of orthologuous genes (COG) database[d] and sequence similarity searches[e]. As expected for small genomes, most of the missing proteins in mycoplasmas and chlamydias are metabolic enzymes that should have no effect on genome rearrangements. Among the proteins involved in translation or transcription, only one kind of transcriptional regulator (COG0789) is missing exclusively in mycoplasmas and chlamydias. It seems, however, that these regulators are not implicated in genome rearrangements. On the other hand, some proteins are missing that are involved in DNA replication and repair[f–j], namely RecG (missing in mycoplasmas and chlamydias), PriA, RuvC and XerCD (only missing in mycoplasmas). Although little is known about the mechanisms for the inversion of large genomic fragments, the lack of these genes might contribute to the absence of inversions in mycoplasmas. Large inversions are often connected with replication (Fig. 1), and mycoplasmas are the only species lacking both RuvC and RecG, which are otherwise present in all other eubacteria studied here (chlamydias have no RecG). This pattern indicates that RecG might be important in genome rearrangement processes.

Although archaea do not have clear orthologs of RecG and some other proteins involved in recombination, this is not surprising because repair enzymes are even different in eubacteria[b], and there might be some proteins with similar functions in archaea. Indeed, in archaea, there are at least two proteins, Hjc (Ref. k) and Hje (Ref. l), neither of which have significant sequence similarity with RuvC, that catalyze the resolution of a Holliday junction. Moreover, in spite of the considerable difference in DNA replication mechanisms between archaea and eubacteria[m], in pyrococci, the mode of replication is similar to that of eubacteria; that is, the replication begins at a defined single origin and proceeds bidirectionally[n]. On the basis of these functional similarities, together with our observations, we speculate that the proteins involved in replication and recombination might also be involved in genome rearrangement in archaea.

References

a Gelfand, M.S. and Koonin, E.V. (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25, 2430–2439

b Eisen, J.A. and Hanawalt, P.C. (1999) A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* 435, 171–213

c Bruno, W.J. *et al.* (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* 17, 189–197

d 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

e Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

f Cox, M.M. *et al.* (2000). The importance of repairing stalled replication forks. *Nature* 404, 37–41

g Liu, J. and Marians, K.J. (1999) PriA-directed assembly of a primosome on D loop DNA. *J. Biol. Chem.* 274, 25033–25041

h Sharples, G.J. *et al.* (1999) Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. *J. Bacteriol.* 181, 5543–5550

i McGlynn, P. *et al.* (2000) Characterisation of the catalytically active form of RecG helicase. *Nucleic Acids Res.* 28, 2324–2332

j Neilson, L. *et al.* (1999) Site-specific recombination at *dif* by *Haemophilus influenzae* XerC. *Mol. Microbiol.* 31, 915–926

k Komori, K. *et al.* (1999) A Holliday junction resolvase from *Pyrococcus furiosus*: functional similarity to *Escherichia coli* RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8873–8878

l Kvaratskhelia, M. and White, M.F. (2000) Two Holliday junction resolving enzymes in *Sulfolobus solfataricus*. *J. Mol. Biol.* 297, 923–932

m Leipe, D.D. *et al.* (1999) Did DNA replication evolve twice independently? *Nucleic Acids Res.* 27, 3389–3401

n Myllykallio, H. *et al.* (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 288, 2212–2215

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.

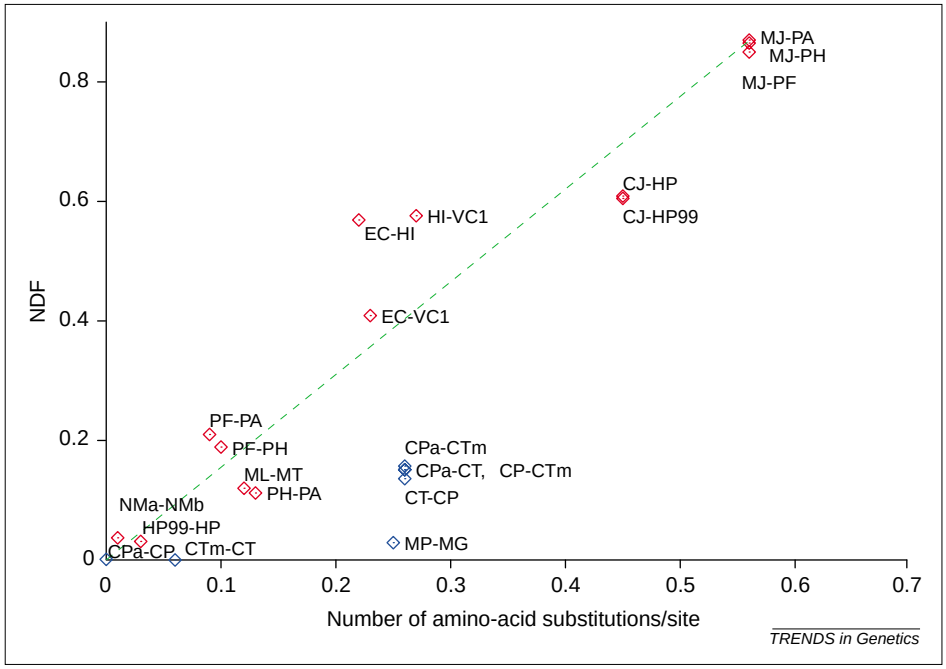

Fig. 2. Distance-dependent genome shuffling. The horizontal axis denotes the averaged number of amino acid substitutions per site for the orthologous genes shared among the 20 genomes analyzed in this study. The vertical axis indicates the neighborhood disruption frequency (NDF; number of orthologous gene neighborhood breakpoints per number of orthologs between a pair of genomes). See legend of Fig. 1 for abbreviations of the genomes. The outliers are indicated in blue. The rest of the points, which are used to draw the regression line (green dotted line), are shown in red. Correlation coefficient values calculated for the data points with and without outliers are $r=0.877$ ($P=1.9\times10^{-7}$) and $r=0.957$ ($P=8.6\times10^{-8}$), respectively.

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[18].

Acknowledgements

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

References

1 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856
2 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
3 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
4 Huynen, M. A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* 54, 345–379
5 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
6 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
7 Louarn, J-M. *et al.* (1991) Analysis and possible role of hyperrecombination in the termination region of the *Escherichia coli* chromosome. *J. Bacteriol.* 173, 5097–5104
8 Read, T.D. *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28, 1397–1406
9 Myllykallio, H. *et al.* (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 288, 2212–2215
10 Sanderson, K.E. and Liu, S-L. (1998) Chromosomal rearrangements in enteric bacteria. *Electrophoresis* 19, 569–572
11 Blanchette, M. *et al.* (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* 49, 193–203
12 Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13, 660–665
13 Lobry, J.R. (1996) Origin of replication of *Mycoplasma genitalium*. *Science* 272, 745–746
14 McLean, M.J. *et al.* (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* 47, 691–696
15 Salzberg, S. L. *et al.* (1998) Skewed oligomers and origins of replication. *Gene* 217, 57–67
16 Qin, M-H. *et al.* (1999) Characterization of the functional replication origin of *Mycobacterium tuberculosis*. *Gene* 233, 121–130
17 Romero, H. *et al.* (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.* 28, 2084–2090
18 Tiller, E.R.M. and Collins, R.A. (2000) Genome rearrangement by replication- directed transcription. *Nat. Genet.* 195–197

M. Suyama† , P. Bork*†‡
†EMBL, Meyerhofstr. 1, D-69012 Heidelberg, Germany.
‡Max Delbrück Center for Molecular Medicine, Berlin–Buch, Germany.
*e-mail: bork@embl-heidelberg.de

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 Boncinelli 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[1]. What are the factors determining cell polarity? In *Caenorhabditis elegans*, *CHE14* encoding a twelve-pass transmembrane protein on the apical surface is required for epithelialization (Michel Labouesse, IGBMC, Strasbourg, France). The protein is required for exocytosis and shares similarity with the *Drosophila* Dispatched protein (better known for its role in Hedgehog secretion), which possesses a lipid-sensing domain perhaps required for targetting to the membrane. The worm surprisingly lacks

  PII: S0168-9525(00)02164-8