Inference and Analysis of the Relative Stability of Bacterial Chromosomes

Eduardo P. C. Rocha
Unite´ Génétique des Génomes Bactériens, Institut Pasteur, Paris, France and Atelier de BioInformatique, Université Pierre et Marie Curie (Paris VI), Paris, France

The stability of genomes is highly variable, both in terms of gene content and gene order. Here I calibrate the loss of gene order conservation (GOC) through time by fitting a simple probabilistic model on pairwise comparisons involving 126 bacterial genomes. The model computes the probability of separation of pairs of contiguous genes per unit of time and fits the data better than previous ones while allowing a mechanistic interpretation for the loss of GOC with time. Although the information on operons is not used in the model, I observe, as expected, that most highly conserved pairs of genes are indeed within operons. However, even the other pairs are much more conserved than expected given the observed experimental rearrangement rates. After 500 Myr, about 50% of the originally contiguous orthologues remain so in the average genome. Hence, the large majority of rearrangements must be deleterious and random genome rearrangements are unlikely to provide for positively selected structural changes. I then use the deviations from the model to define an intrinsic measure of genome stability that allowed the comparison of distantly related genomes and the inference of ancestral states. This shows that clades differ in genome stability, with cyanobacteria being the least stable and $\gamma$-proteobacteria the most stable. Without correction for phylogeny, free-living bacteria are the least stable group of genomes, followed by pathogens, and then endomutualists. However, after correction for phylogenetic inertia (or the removal of cyanobacteria from the analysis), there is no significant association between genome stability and lifestyle or genome size. Hence, although this method has allowed uncovering some of mechanisms leading to rearrangements, we still ignore the forces that differentially shape selection upon genome stability in different species.

Introduction

The stability of genomes results from a mutation-selection balance. As a result of different selection pressures, effective population sizes, and rearrangement rates, some genomes are significantly more stable than others. Different selective forces have been recently uncovered that shape gene order both in prokaryotes and eukaryotes (Hurst, Pal, and Lercher 2004; Rocha 2004b). Many of the selective elements shaping the organization of bacterial chromosomes result from fundamental processes of life such as gene expression and replication (reviewed in Lawrence 2003; Bentley and Parkhill 2004; Rocha 2004c). Gene expression shapes the chromosome organization by selecting the clustering of functionally related genes into operons (Jacob and Monod 1961) and supraoperons (Lathe, Snel, and Bork 2000). This allows the optimization of gene expression regulation (Hershberg, Yeger-Lotem, and Margalit 2005). Replication asymmetries result in the preferential positioning of genes, especially essential genes, in the leading strand (Rocha and Danchin 2003), highly expressed genes near the origin of replication in fast-growing bacteria (Chandler and Pritchard 1975), and balanced horizontal gene transfer, and possibly gene deletion, among the two bacterial repli- chores (Bergthorsson and Ochman 1998). These organizational features allow bacteria to be substantially fitter, and their disruption is bound to be deleterious.

Inversions have experimentally been found to have very different fitness effects depending on the exact rearrangements and the selective features that are disrupted. Many cases have been reported of significantly lower growth rates and even lethality resulting from inversions (Hill and Gray 1988; Rebollo, François, and Louarn 1988; Segall, Mahan, and Roth 1988; Campo et al. 2004). Rearrangements are thought to occur mostly by two types of mechanisms both involving intrachromosomal recombination (reviewed in Hughes 1999; Rocha 2004b). Homologous recombination can target inverted repeated elements, such as rDNA operons or insertion sequences (IS), and lead to large inversions (Hill and Harnish 1981; Nakagawa et al. 2003; Parkhill et al. 2003). IS and other elements, for example, phages, may additionally lead to chromosome rearrangements because of the activity of their recombinases (Gray 2000). One should note that rearrangements in Escherichia coli are very frequent in the laboratory (Hill and Harnish 1981), but few are found to have been fixed since the divergence from Salmonella enterica, about ~100 MYA. This is usually thought to be the result of natural selection acting toward eliminating these deleterious events, which assumes that most inversions are at least slightly deleterious. Hence, the frequency of rearrangement events and the subsequent purging of deleterious ones by natural selection will determine the long-term conservation of gene order in bacterial genomes. The calibration of bacterial stability, understood as the inverse of the frequency of gene order disruption, can thus shed some light on both rearrangement mechanisms and selective processes shaping the organization of the chromosome.

The analysis of complete bacterial genomes has revealed that the obligatory intracellular mutualists such as Buchnera are among the most stable genomes. In these sexually isolated species with small effective population sizes, horizontal transfer cannot compensate for gene loss. Combined with relaxed selection this leads to small genomes with few repeated elements, no IS, and few recombination genes (Shigenobu et al. 2000; Lawrence, Hendrix, and Casiens 2001; Dale et al. 2003; Silva, Latorre, and Moya 2003). These genomes are then extremely stable, with no large rearrangement apparent for over 100 Myr (Tamas et al. 2002), just some gene deletions, possibly by illegitimate recombination between small repeats (Mira, Ochman, and Moran 2001; Silva, Latorre, and Moya 2001; Rocha 2003b). A more nuanced picture is given by the genome...
of pathogenic obligatory intracellular bacteria such as the ones of \textit{chlamydia} and \textit{rickettsia}, which are nevertheless quite stable and nearly devoid of very large repeats (Kalman et al. 1999; Frank, Amiri, and Andersson 2002). Intermediate genome stability is found in \textit{E. coli} and \textit{S. enterica}. These genomes show abundant horizontal gene transfer, and some close strains differ by more than 25% of their genetic material (Berghorsson and Ochman 1998; Ochman and Jones 2000). However, this has little influence on the overall genome structure because only one large rearrangement differentiates \textit{E. coli} K12 from \textit{S. enterica typhimurium} (McClelland et al. 2001). Among the least stable genomes, one finds pathogens such as \textit{Bordetella pertussis} and \textit{Yersinia pestis} that have very recently diversified from a more stable genome by losing some genes and by gaining a large number of IS. These genomes show a large number of recent rearrangements (Parkhill et al. 2003; Chain et al. 2004).

It must be emphasized that most characterizations of genome stability have relied on the comparisons of a small number of closely related genomes (Eisen et al. 2000; Tillier and Collins 2000; Suyama and Bork 2001; Parkhill et al. 2003). This poses a problem if one wants to define a calibrated measure of stability, which would allow to compare very distant genomes and test the association of stability with ecological and population parameters. There are currently two major methodological approaches to the study of gene order. One approach aims at determining the rearrangement distance between two genomes. This distance is an estimation of the number of rearrangement events that took place since speciation and allows the analysis of evolutionary scenarios (Sankoff 2003). However, it is not yet adapted for the analysis of large sets of bacterial genomes at large rearrangement distances. As a result, most large-scale analyses of gene order evolution in bacteria have used empirical measures of distance based on pairwise comparisons of gene order (Huynen and Bork 1998; Tamames 2001; Rocha 2003a). Here I try to calibrate bacterial genome stability by building an index of genome stability that takes into account the average loss of gene order through time. A nonlinear regression fits the available data, and this allows the assessment of the relative stability of each genome. I then explore the potential association of the stability of genomes with their phylogenetic grouping, genome size, and lifestyle.

\textbf{Methods and Data}

\textbf{Methods}

\textbf{Data}

A total of 126 genomes from six clades with more than eight available genomes were analyzed: Actinobacteria, Cytophagales, Firmicutes, \textgamma-proteobacteria, \textbeta-proteobacteria, and \textalpha-proteobacteria (Supplementary Table 1, Supplementary Material online). Genomes were taken from GenBank. The classification of bacteria according to their lifestyle was taken from the literature.

\textit{Assignment of Orthology and Phylogenetic Analysis}

Orthologues were defined as unique reciprocal best hits (using the score of a global alignment where the gaps on the edges of the smallest sequence are not penalized) with at least 40% similarity in amino acid sequence and less than 20\% of difference in protein length. The analysis of orthology was made for every pair of genomes within each phylogenetic group. Because of gene transfer, gene deletion, and sequence divergence, the number of orthologues decreases with the evolutionary distance. I tested if the most distant comparisons in each clade still included a significant number of genes. Among \textgamma-proteobacteria, the comparison between \textit{E. coli} K12 (~4,000 genes) and \textit{Xylella fastidiosa} Temecula1 (~2,000 genes) included over 1,000 orthologues. Among Firmicutes, the comparison of \textit{Bacillus subtilis} (~4,100 genes) with \textit{Clostridium perfringens} (~2,800 genes) included over 1,100 orthologues. Hence, the number of orthologues involved in the comparisons is always very large. The evolutionary distances between bacteria were computed from the 16S rDNA subunit, with Tree-Puzzle (Schmidt et al. 2002) using the Haseawa, Kishino, and Yano (HKY) + \Gamma model. The assessment of phylogenetic inertia was done using CONTINUOUS (Pagel 1999) and the inference of ancestral states with COMPARE (Martins and Hansen 1997), both using the 16S rDNA phylogenetic tree as a reference.

\textit{Operons}

Many recent methods have been developed to unravel the operon structure of genomes, but most use gene order conservation (GOC) to attain higher accuracy. Because I am interested in gene conservation and on the constraints imposed on it by operons, the use of such methods would introduce some circularity in the analysis. I have thus identified operons in \textit{B. subtilis} and \textit{E. coli} based on gene transcription sense, rho-independent terminators identified by TransTerm (Ermolaeva et al. 2000), and intergenic distance (Salgado et al. 2000).

\textit{GOC}

We are interested in analyzing the changes in gene order when horizontal gene transfer is excluded (as in Huynen and Bork 1998; Tamames 2001; Rocha 2003a). For this I make pairwise comparisons starting from the list of orthologues between the genomes. In such pairwise comparisons, the GOC is defined as the relative frequency with which two contiguous genes in a genome have their respective orthologues also contiguous in the ordered list of orthologues of the other genome. Hence, GOC can be estimated between a pair of genomes as the number of pairs of orthologues that are contiguous in the two genomes divided by the total number of orthologues:

\begin{equation}
\text{GOC} = \frac{N_{\text{orthologues contiguous}}}{N_{\text{orthologues}}}.
\end{equation}

This condition of contiguity is somewhat stringent, but simulations have shown that in bacteria the addition of a neighborhood (i.e., a vicinity distance larger than one) changes very little of the results and unnecessarily complicates the statistical analysis (Rocha 2003a). The analysis of eukaryotes, because of gene loss following genome duplication, will certainly require the inclusion of a neighborhood. When the number of orthologues is large, as in this work, the probability of a pair of genes being contiguous in the two genomes by chance is very small.
Models of GOC

Several empirical models have been proposed to fit the loss of GOC with time ($t$), with a parameter $\alpha$ to be adjusted by regression. Tamames proposed a sigmoid loss of GOC with time (Tamames 2001):

$$GOC = \frac{2}{1 + e^{-\alpha t}}.$$ (model 0)

I previously proposed a square-root dependence (Rocha 2003a):

$$GOC = 1 - \sqrt{\alpha t}.$$ (model 1)

Although this equation fitted best the data on $\gamma$-proteobacteria, it has the obvious problem that for large enough $t$ GOC will take negative values, which it cannot. A reciprocal transformation has also been proposed (Suyama and Bork 2001):

$$\frac{1}{GOC} = \alpha t + 1.$$ (model 2)

All these models were purely empirical, that is, no mechanistic model was proposed to explain their fit. Because model 2 fits well the data, I tried to infer the dynamic behind it. In fact, $1/GOC = \alpha t + 1$ arises from solving the following differential equation:

$$\frac{dGOC}{dt} = -\alpha GOC^2.$$

This is equivalent to consider a decrease of GOC with time that is negatively proportional to the square of the GOC at a given moment. A different formulation of the model is:

$$GOC = \frac{1}{1 + \alpha t}.$$ (model 3)

I also derived a probabilistic model, inspired from Wolf et al. (2001), that allows a deeper insight on the mechanisms leading to genome rearrangements. A genome is an ordered circular (or linear) sequence of genes that can change by rearrangement (fig. 1). If the probability of two contiguous genes staying together in consecutive generations is $p$, and is constant throughout the genome and through time, then the likeness of two genes staying contiguous after $t$ generations is given by $P = p^t$. Assuming that the probability of splitting contiguous genes is the same for all genes, GOC is an estimator of $P$. Hence model 3 is:

$$GOC = p^t.$$ (model 3)

Because some gene pairs are under stronger selection than others, it would be better to consider a model where the gene pairs are divided into two classes of $n_1$ fast- and $n_2$ slow-rearranging pairs ($p_f$ and $p_s$, $n_1 + n_2 = N$) (fig. 1), then GOC can be modeled by:

$$GOC = P_f + P_s = \frac{(n_1 p_f + n_2 p_s)}{N}.$$ (model 4)

Results and Discussion

The Change of GOC Through Time

I defined GOC between two genomes as the frequency of pairs of orthologous genes that are contiguous in both genomes. Hence, GOC varies between one (complete conservation of gene order) and zero (no conservation). High values of GOC can be obtained if the genomes are very stable or if they have diverged very recently. As a result, if one is interested in understanding the patterns of genome stability, one must start by modeling the time dependence of GOC. The loss of GOC with time, that is, the way GOC decays when the comparisons are made between increasingly divergent genomes, has been modeled in several ways (see Methods and Data for the precise description of the models). An early analysis suggested that GOC decreased with time following a sigmoid transformation (model 0), which would reflect a cooperative loss of gene order with time (Tamames 2001). There is little evidence supporting that claim from this much larger data set of 126 complete genomes from six clades and when the divergence of the 16S rRNA subunit is used as a proxy for time (fig. 2). I have previously suggested that in $\gamma$-proteobacteria GOC decreased with the square root of time (model 1) (Rocha 2003a). Although this empirical transformation fits well...
the data on close comparisons, it performs poorly when distant ones are included (fig. 2). Finally, it has also been proposed that 1/GOC varies linearly with time (model 2) (Suyama and Bork 2001). This model is better than the previous models, especially for more distant comparisons (akaike information criterion [AIC], P < 0.001).

A probabilistic-based model accounts for the probability of contiguous gene pairs being conserved per unit of time (\( p \)), and leads to GOC = \( p^t \) (model 3). This model assumes that all pairs of contiguous genes have the same probability of being separated per unit of time. This assumption is unreasonable because the linkage of pairs of contiguous genes is most certainly under varying selective pressures. As a result of this limitation, the regression of GOC with time using this model performs poorly for the most divergent comparisons (fig. 2). If, however, one allows for two equally sized classes of contiguous pairs of genes, one obtains a model (model 4) GOC = \( \frac{p_f^t + p_s^t}{2} \), which fits very well the data, explaining 73% of the variance (\( P < 0.001 \)). One class of pairs of contiguous genes corresponds to slow-separating pairs (\( p_s = 0.1246 \)) and the other to fast-separating pairs (\( p_f = 2.8 \times 10^{-10} \), significantly different from 0, \( P < 0.001, F \) test) (fig. 2). The fit of this model is the best (\( P \ll 0.001, AIC \) tests) and immediately suggests that selection on the conservation of operons could be a major force behind GOC. Operons are on average slightly more than three genes long and include two-thirds of the genome (Zheng et al. 2002). Hence, roughly half of the contiguous genes belong to the same operon and half to different operons. This is consistent to the hypothesis that half of the pairs separate fast and half separate slowly.

In spite of their intrinsic differences, both model 2 and model 4 fit reasonably well the data. Model 4 assumes the existence of two classes of pairs of contiguous orthologues, one splitting apart faster than the other, whereas model 2 assumes a continuum of increasing selective pressures on

GOC between pairs of contiguous genes (see Methods and Data). Model 4 fits better the data and one would expect the two classes of pairs of orthologues to concern pairs within operons (\( p_o \)) versus pairs between contiguous operons (\( p_i \)). To test this hypothesis, I identified operons in B. subtilis and E. coli and computed separately the change of GOC with time for contiguous genes in the same operon and for contiguous genes in contiguous operons. Not only GOC is much more conserved for genes in operons but the results are also quantitatively close to the ones expected by the previous fit of model 4 (fig. 3). Hence, the selection acting upon operons is the major force acting against the disruption of gene order and is responsible for the particularly low level of rearrangements for pairs of orthologous genes.
within an operon. If one calibrates the evolutionary distance using 100 Myr for the *Escherichia/Salmonella* divergence (Ochman and Wilson 1987), one finds that bacteria separated by over 500 Myr still have conserved the contiguity of \(~80\%\) of the within-operon gene pairs.

The above mentioned \(p_1\) estimation of \(2.8 \times 10^{-10}\) may seem very small and suggests a very high rate of rearrangements for these pairs. However, one should bear in mind that experimental data suggest that in the absence of selection, rearrangement rates can be on the order of \(10^{-4}/\text{generation}\) among rDNA operons of *E. coli* K12 (Hill and Harnish 1981). This rate is an underestimation of the real rearrangement rate per generation per genome because rDNA elements are not the only repeated elements that are sources of inversions and because strongly deleterious rearrangements are probably missed in the analysis (e.g., if they are lethal). Assuming 100 Myr for the separation between *Escherichia* and *Salmonella* and 200 generations/year for these bacteria, then \(p_1 = 2.8 \times 10^{-10}\) translates in a genomic frequency of rearrangement per generation of \(10^{-7}\), which is still much less than the conservative frequency of \(10^{-5}\) observed in the laboratory. This strongly suggests that even among this class, the vast majority of rearrangements are deleterious. Given the large effective sizes of bacterial populations, slightly deleterious effects are probably enough to purge out such rearrangements from natural populations.

### Relative Stability of Bacterial Genomes

Because model 4 is the one presenting the best fit to the data, it was the one used throughout the rest of the work. Each genome participates in a subset of the pairwise comparisons used to make the regression analysis. The stability of each genome is then defined as the average of the residuals resulting from the nonlinear regression of model 4 for the comparisons where the genome participates. Hence, the stability of a genome, for example, *B. subtilis*, is the average of the residuals for all pairwise comparisons where *B. subtilis* participates. The stability of each genome is then the average deviation of the observed values of GOC in the pairwise comparisons from model 4 to the expected values given a certain phylogenetic distance (Supplementary Table 1. Supplementary Material online). Naturally, the estimates of stability are more precise for the groups containing the larger number of genomes (i.e., *Firmicutes* and \(\gamma\)-proteobacteria). The stability thus calculated matches previous reports for simple pairwise comparisons between genomes. This is the case for genomes that are less stable such as the ones of *Streptomyces* (Vollf and Altenbuchner 1998), *Bacillus cereus* (Kolsto 1997), *Wolbachia* (Wu et al. 2004), or *Neisseria meningitidis* (Dempsey, Wallace, and Cannon 1995). Some other particularly unstable genomes within each clade include *Streptococcus pneumoniae*, *Enterococcus faecalis*, *Synechocystis*, and *Thermosynechocystis*. On the other extreme, this analysis fits previous observations that the genomes of \(\gamma\)-proteobacteria obligate endomutualists such as *Buchnera* and *Candidatus* are extremely stable. Also stable, but less so, are the genomes of some obligate intracellular pathogens, for example, *Tropheryma* and *Mycobacterium leprae*, although other intracellular obligatory parasites such as *rickettsia* are not particularly stable.

As expected, closely related strains or species tend to have similar stability indices. This is because, contrary to most analyses, this one takes into account a larger amount of comparisons at deeper evolutionary distances. Hence, very recent changes in stability, such as the high number of rearrangements in *Shigella* strains within *E. coli* (Jin et al. 2002), have little weight in the overall analysis. To test if the more recent evolutionary history is associated with a change in stability, I computed two separate measures of stability, one including only the 50% closer comparisons and another including only the 50% most distant comparisons. I then tested the difference between the two values and selected the genomes for which recent and older comparisons are significantly different \((P < 0.05, t\text{-tests with sequential Bonferroni correction for multiple tests within the clades})\). Naturally, the significance of differences also depends on the number of genomes of each group, and it is not surprising that the differences are not significant among the less sampled clades. Yet, this analysis points to significant differences in some genomes among the most represented groups. Notably, it confirms that obligatory endomutualists of \(\gamma\)-proteobacteria have become more stable recently (*Buchnera*, *Wigglesworthia*, and *Buchnera*). Inversely, instability is recent among *Actinobacter*, *S. pneunmiae*, *Legionella pneumophila*, *Pseudomonas*, *Xanthomonas axonopodis*, *Wolbachia*, *Brucella melitensis*, and *E. faecalis*. It is remarkable that many of the latter bacteria are thought to be recent pathogens. When sufficient genomes become available, this method will allow following the evolutionary change of stability through time, by comparing a genome with increasingly proximal ones. Using the current data set, such differences are already visible among \(\gamma\)-proteobacteria and *Firmicutes*.

Genomes may be related by common selective constraints acting upon them. However, genomes also share an evolutionary history. Hence, closely related genomes are expected to be similar both because they tend to have more of such selective constraints in common and because they shared a common history until a more recent period of time. To distinguish selective explanations for the stability of genomes from such phylogenetic inertia (see *Association of Stability with Lifestyle*), one can use the comparative method (Felsenstein 1985). For this I first tried to understand the evolution of the stability through time taking into account the 16S rDNA phylogeny. Using CONTINUOUS (Pagel 1999), I checked that this variable evolved according to a Brownian model \((\lambda = 0.99, \text{not significantly different from 1, } P > 0.05)\). Hence, one can use standard methods to infer the ancestral states for the stability of the genomes. This was done using COMPARE (Martins and Hansen 1997) and allowed tracking the changes in stability along the evolutionary history of these bacteria (fig. 4).

The results displayed in figure 4 suggest that some clades are more stable than others. This could be confirmed by pooling the stability of genomes for each clade (fig. 5, Kruskal-Wallis test, \(P < 0.001\)). To check which pairs of clades are different, a Tukey-Kramer HSD test was used which revealed the existence of three groups \((P < 0.01)\). The most stable clade is composed of \(\gamma\)-proteobacteria partly, but not entirely, due to the presence of all obligatory endomutualists in this group. Then follows the group...
including the other proteobacteria, the Firmicutes, and the Actinobacteria. Cyanobacteria is the group containing the least stable genomes. There are a couple of observations that reinforce this classification: cyanobacteria tend to have little, if any, replication-associated biases (Rocha 2004a), and site-specific rearrangement systems are known to operate in several of these genomes (Carrasco and Golden 1995). This result, together with the analysis of the ancestral states, strongly suggests that rearrangements are fixed at very different rates in different clades.

**Association of Stability with Lifestyle**

Lifestyle has often been associated with genome stability, with endomutualists being regarded as stable (Tamas et al. 2002) and pathogens as unstable (Hughes 2000; Hacker, Hentschel, and Dobrindt 2003). This includes extracellular recent pathogens such as *Y. pestis* and *Bordetella* (Deng et al. 2002; Parkhill et al. 2003; Chain et al. 2004). I tried to test these associations by dividing the genomes into free-living bacteria (including facultative commensals), facultative pathogens, obligatory pathogens, and obligatory endomutualists (fig. 6). This analysis showed significant differences between the groups (*P* < 0.001, Kruskal-Wallis), and multiple comparisons indicate that endomutualists form a significantly more stable group, whereas free-living bacteria are significantly less stable (*P* < 0.01, Tukey-Kramer HSD). Pathogenic bacteria have an average stability. Endomutualists and obligatory pathogens have small genomes and often have few repeats, which suggest an inverse correlation between genome size and stability. Yet, this correlation is very close to zero (Spearman *ρ* = −0.02, *P* > 0.5), indicating that genome stability is independent of genome size. Hence, the stability of these genomes is more likely to result from the lack of the agents that more often cause recombination such as IS, repeats, and homologous recombination.

The phylogenetic correlation between the species complicates the analysis of the association of lifestyle with genome stability. For example, all cyanobacteria are free-living and it is here found that cyanobacteria are less stable. So, one should assign the instability of this group to their
lifestyle or to their clade? More importantly, the difference of stability between species is dependent on their common evolutionary history. This means that the association between two variables (e.g., lifestyle and stability) may be a fortuitous consequence of phylogenetic inertia (Felsenstein 1985). To separate the hypothesis that the two variables are intrinsically correlated or that their correlation is due to phylogenetic inertia we used the comparative method (Pagel 1999). One of the classes, obligatory endomutualists, is a monophyletic clade of γ-proteobacteria. Hence, in this case, the test cannot be performed and one cannot decide between the two hypotheses. Because both classes of pathogens showed similar average stability, we pooled them together and tested if this group is significantly more stable than the free-living bacteria. Because stability is a continuous variable and lifestyle a categorical variable with two states, CONTINUOUS was used to evaluate the robustness of this association (Pagel 1999). The contrast analysis indicated a much smaller and nonsignificant correlation between stability and lifestyle ($r = 0.04$) than in the original data ($r = 0.33$). This suggests that phylogenetic inertia may explain the largest part of the effect of lifestyle on genome stability. Hence, although it seems clear that pathogens are not less stable than free-living bacteria, one cannot at the moment exclude phylogenetic inertia as the reason for the inverse association.

**Conclusion**

The definition of an index of genome stability allows testing evolutionary hypothesis concerning the occurrence and fixation of structural changes in the bacterial chromosome. When more close eukaryotic genomes become available, one may also use it to understand the patterns of eukaryotic rearrangements and to what extent is gene order under selection in these genomes. However, the analysis will have to be adapted to tackle the much more frequent events of gene and even genome duplication. The method will also allow, when more data become available, to study the change in stability through time in a lineage by using different thresholds of phylogenetic distance when choosing the genomes from which to compute stability. Here, I showed that one could identify some cases of recent increased genome stability and instability. Interestingly, bacteria that most recently became more stable are obligatory endomutualists, and bacteria that have recently become less stable are for the majority recent pathogens. Several of these observations match previous independent analysis. A major hurdle that remains in this analysis concerns the use of a given phylogeny. This can be circumvented if the use of the comparative method includes phylogenetic uncertainty (Huelsenbeck, Rannala, and Masly 2000). However, any such study will still assume that the substitution rates are the same in the 16S of the different lineages. If this is not true, then deviations from the model can be attributed to differences in genome stability or to differences in the evolutionary rate of the phylogenetic markers. The data on the analysis of closely related genomes (e.g., among enterobacteria) suggest that the stability of genomes evolves faster than the evolutionary rate of sequences. However, the effect of deviations to the molecular clock should be taken into account in future studies.

The first comparisons between distant genomes revealed little conservation of gene order (Mushegian and Koonin 1996; Huynen and Bork 1998; Itoh et al. 1999). This was surprising because decades of research on the regulation of gene expression suggested that this was under strong selection. Further, several experimental works measured the impact of large rearrangements on bacterial fitness (Louarn et al. 1985; Hill and Gray 1988; Campo et al. 2004). However, detailed comparisons of operon composition have shown that a significant number of operons are partially conserved between the distantly related *E. coli* and *B. subtilis* (Overbeek et al. 1999; de Daruvar, Collado-Vides, and Valencia 2002). This brings to the fore the problem I tried to tackle in this paper: the stability of genomes must be interpreted after the comparisons have been calibrated by the time since divergence. When this is done and one compares genomes that have diverged over 500 MYA, one still observes a very significant conservation of gene order. This long-range conservation seems to stem essentially from the existence of operons and explains why no similar conservation was found in the comparison of distantly related yeasts (Dujon et al. 2004). However, even pairs of genes in contiguous operons seem to be separated much slower than expected given the rearrangement rates observed in the laboratory. This may partly be because of breakpoint reuse in large repeated elements, such as tRNA operons, or if laboratory conditions, by some unknown reason, overestimate the rate of chromosome rearrangements. However, supraoperonic gene order has been observed in bacterial genomes, possibly to optimize the use of regulatory regions (Korbel et al. 2004; Warren and ten Wolde 2004; Hershberg, Yeger-Lotem, and Margalit 2005), and it has been suggested that contiguity between operons can be as much under selection as contiguity between genes within an operon. Hence, it seems reasonable to suggest that the GOC data reflects such selective constraints.

Previous analysis, with fewer data, suggested that gene order is randomized (except for gene clusters with functional constraints such as operons) if the 16S rRNA distance exceeds 10% (Huynen and Bork 1998). The current work confirms that this is true on average, but that for unstable
E. coli observed number of rearrangements on the order of a dozen the result of conflicting forces. On one hand, rearrangement is small, at best. In fact, stability is likely to be surprising that free-living bacteria are the least stable. How-

different challenges imposed on the species, it is not sur-

endomutualists are more stable. If stability is the by-

pathogenicity itself. Free-living bacteria are less stable and

relation with a recent change in ecological niche than with

(Bentley and Parkhill 2004). Comparative genomic struc-


Berghorsson, U., and H. Ochman. 1998. Distribution of chromo-

some length variation in natural isolates of Escherichia coli. Mol.


Campo, N., M. J. Dias, M. L. Duveran-Mingot, P. Ritzenthaler,


Supplementary Material

Supplementary Table 1 is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I am most grateful to the people with whom I have discussed the issues presented in this article and especially to Guillaume Achaz and Isabelle Gonçalves for comments and criticisms on earlier versions of this text.

Literature Cited


Rocha, 2003a). The num-

ber of rearrangements is expected to be higher in genomes that are larger, because they have more duplicated elements (Achaz et al. 2002), or less sexually isolated, because these genomes will be more frequently invaded by IS, phages, and restriction and modification systems. However, larger bacterial genomes are probably also associated with higher effective population sizes because they correspond to commensals and free-living bacteria. These compensating forces might explain why there is no general correlation between genome size and stability or between lifestyle and stability when phylogenetic inertia is taken into account. Obligatory symbionts, either pathogens or mutualists, have lower effective population sizes, but also fewer repeats, few horizontal gene transfers, and they often miss genes associated with homologous recombination. They are thus more stable not because of more efficient selection, but for lack of elements capable of disrupting the genome.

The results of this work indicate that there is no simple clear-cut association between genome stability and lifestyle as both very unstable and very stable genomes are included in each class. Although it has often been suggested that pathogens are less stable because of adaptation to the host (Hughes 1999; Hacker, Hentschel, and Dobrindt 2003), I found that they have an intermediate stability. The excep-
tion may be some recent pathogens that were found to have become less stable recently. Hence, instability may be more related with a recent change in ecological niche than with pathogenicity itself. Free-living bacteria are less stable and endomutualists are more stable. If stability is the by-product of the need to adapt and this is the reflex of the number of different challenges imposed on the species, it is not surprising that free-living bacteria are the least stable. How-

ever, given the strong purifying selection acting on rearrangements, the possibility of a positive effect of a rear-

arrangement is small, at best. In fact, stability is likely to be the result of conflicting forces. On one hand, rearrange-
maments are expected to be deleterious, otherwise they would be fixed more frequently. Hence, more efficient selection, for example, due to larger effective population sizes, will tend to increase the stability because most rearrangements will be purged from the population. This may explain why the bottleneck-prone recent pathogens have recently be-
come less stable. On the other hand, the occurrence of rearrangements is strongly dependent on substrates for recombination events, such as repeats. This is why highly repeated genomes are less stable (Rocha 2003a). The num-

ber of rearrangements is expected to be higher in genomes that are larger, because they have more duplicated elements (Achaz et al. 2002), or less sexually isolated, because these genomes will be more frequently invaded by IS, phages, and restriction and modification systems. However, larger bacterial genomes are probably also associated with higher effective population sizes because they correspond to commensals and free-living bacteria. These compensating forces might explain why there is no general correlation between genome size and stability or between lifestyle and stability when phylogenetic inertia is taken into account. Obligatory symbionts, either pathogens or mutualists, have lower effective population sizes, but also fewer repeats, few horizontal gene transfers, and they often miss genes associated with homologous recombination. They are thus more stable not because of more efficient selection, but for lack of elements capable of disrupting the genome.

Supplementary Material

Supplementary Table 1 is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I am most grateful to the people with whom I have discussed the issues presented in this article and especially to Guillaume Achaz and Isabelle Gonçalves for comments and criticisms on earlier versions of this text.

Literature Cited


Takashi Gojobori, Associate Editor

Accepted November 2, 2005