Measuring genome evolution
(ortholog/synteny/computer analysis/horizontal gene transfer)

MARTIN A. HUYEN* AND PEER BORK

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany, and Max-Delbrück-Centrum for Molecular Medicine, 13122 Berlin-Buch, Germany

ABSTRACT The determination of complete genome sequences provides us with an opportunity to describe and analyze evolution at the comprehensive level of genomes. Here we compare nine genomes with respect to their protein coding genes at two levels: (i) we compare genomes as “bags of genes” and measure the fraction of orthologs shared between genomes and (ii) we quantify correlations between genes with respect to their relative positions in genomes. Distances between the genomes are related to their divergence times, measured as the number of amino acid substitutions per site in a set of 34 orthologous genes that are shared among all the genomes compared. We establish a hierarchy of rates at which genomes have changed during evolution. Protein sequence identity is the most conserved, followed by the complement of genes within the genome. Next is the degree of conservation of the order of genes, whereas gene regulation appears to evolve at the highest rate. Finally, we show that some genomes are more highly organized than others: they show a higher degree of the clustering of genes that have orthologs in other genomes.

Molecular evolution usually is studied at the level of single genes. With the determination of genome sequences we have an opportunity to study it at a higher, comprehensive level, that of complete genomes. This leads to the pertinent question: how can genomic information be used to obtain useful information concerning genome evolution? The goal of this paper is to create baseline expectations for measures of genome distances that are based on gene content. By describing some general patterns one also can identify the exceptions. Measuring evolution at the level of complete genomes is pertinent as it is, after all, the principal level for natural selection. Furthermore, it is intermediate to levels at which evolution has long been studied: namely, the molecular level in genes and genotypes, and the organismal level in the fossil record. The genome in principle contains all of the information necessary to bridge the gap between genotype and phenotype. For example, by knowing the functions of the genes in a genome of a species we can postulate a model for its complete metabolism. However, we have to be careful not to overstate our expectations. The situation might turn out to be analogous to that of proteins, for which, in principle, all information necessary to determine three-dimensional structures in the form of amino acid sequences is known, yet we remain unable to predict their tertiary structures.

Genomes can be analyzed and compared for various features: e.g., nucleotide content, compositional biases of leading and lagging strands in replication (e.g., in Escherichia coli) (1), dinucleotide frequencies (2), the occurrence of repeats (e.g., in virulence genes of Haemophilus influenzae; ref. 3), RNA structures, coding densities, protein coding genes, operons, the size distribution of gene families (4), etc. They also can be compared at a variety of levels: a first-order level where we regard the genome as a “bag of genes” without taking account of interactions between the various components, and a second-order level that considers whether properties of genomes are cross-correlated (e.g., the absence of certain polynucleotides together with the presence of restriction enzymes that specifically cut these polynucleotides; ref. 5). In this paper we focus on first- and second-order patterns in protein coding regions in genomes. Specifically we measure: (i) the fraction of orthologous sequences between genomes, (ii) the conservation of gene order between genomes, and (iii) the spatial clustering of genes in one genome that have an ortholog in another genome. We correlate these measures with the divergence time between the genomes compared. It is not our goal to define new distance measures to construct phylogenetic trees. Rather it is to analyze the conservation and differentiation of patterns between genomes, to show how we can extract useful information from these, and to analyze at what relative time scales they change. The analyses are done on the first nine sequenced Archaea and Bacteria that were publicly available: H. influenzae (6), Mycoplasma genitalium (7), Synechocystis sp. PCC 6803 (8), Methanococcus jannaschii (9), Mycoplasma pneumoniae (10), E. coli (1), Methanobacterium thermoaerotrophicum (11), Helicobacter pylori (12), and Bacillus subtilis (13). Although the total number of publicly available genome sequences is growing rapidly, the trends that we observe should remain largely unchanged with the comparison of new species, given the diverse range of evolutionary distances of the species compared in this paper.

Methodological Issues in Comparisons of Genomes

Identification of Orthologous Genes. Defining orthology. In comparing the genes of different genomes it is important that we avoid comparisons of “apples and pears”: i.e., that we are able to identify which genes correspond to each other in the various genomes. Fitch (14) introduced the term “orthologs” for genes whose independent evolution reflects a speciation event rather than a gene duplication event. “Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example, alpha hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact)” (14). Note that orthology and paralogy are
defined only with respect to the phylogeny of the genes and not with respect to function.

**Identifying orthology by using relative levels of sequence identity.** Ideally one would expect that the orthologous genes of two genomes are those that have the highest pairwise identity, having bifurcated relatively recently compared with genes that duplicated before the speciation. The most straightforward approach to identifying orthologous genes is to compare all genes in genomes with each other, and then to select pairs of genes with significant pairwise similarities. A pair of sequences with the highest level of identity then is considered orthologs.

**Auxiliary information for detection of orthology.** Auxiliary information that is useful to assess orthology is “synteny”: the presence in both genomes of neighboring sequences that are also orthologs of each other. As shown below, there is little conservation of the order of genes in genomes in evolution at a time when divergence of their orthologous genes reaches a level of 50% amino acid identity (see Fig. 3). Hence the potential for using synteny for identifying orthologs is limited mainly to genomes that have speciated only relatively recently. A second type of auxiliary information that can be used is the comparison of genes with those of a third genome. If two genes from different genomes have the highest level of identity both to each other and to a single gene from a third genome, then this is a strong indication that they are orthologs (see ref. 15 for a large-scale implementation of this idea). However for a large fraction of genes identifying orthologs by relative sequence identity is hampered by a variety of evolutionary processes. We describe these in the following sections.

**Sequence divergence.** At large evolutionary distances, e.g., between Archaea and Bacteria, sequence similarities may be eroded to such an extent that the distance between orthologous sequences is similar to that between sequences that are merely part of the same gene family. More dramatically, homolog sequences can diverge “beyond recognition,” such that the similarity between two orthologs is not higher than the similarity between sequences that are not part of the same gene family and automatic procedures for the recognition of homology fail. A recent survey of genes in *Drosophila* shows that one-third of the cDNAs code for very fast evolving genes, for which the frequency of amino acid substituting mutations is only a 2-fold lower than that of silent mutations, leading to a situation where homologous proteins are barely recognizable after 8,000 years of evolution (16).

**Orthology in multidomain proteins.** Orthology in multidomain proteins is often more difficult to detect than orthology in single protein domains. A first step toward modularity, the presence in both genomes of neighboring sequences that are also orthologs of each other, then is considered orthologs of each other. As shown below, there is little conservation of the order of genes in genomes in evolution at a time when divergence of their orthologous genes reaches a level of 50% amino acid identity (see Fig. 3). Hence the potential for using synteny for identifying orthologs is limited mainly to genomes that have speciated only relatively recently. A second type of auxiliary information that can be used is the comparison of genes with those of a third genome. If two genes from different genomes have the highest level of identity both to each other and to a single gene from a third genome, then this is a strong indication that they are orthologs (see ref. 15 for a large-scale implementation of this idea). However for a large fraction of genes identifying orthologs by relative sequence identity is hampered by a variety of evolutionary processes. We describe these in the following sections.

**Sequence divergence.** At large evolutionary distances, e.g., between Archaea and Bacteria, sequence similarities may be eroded to such an extent that the distance between orthologous sequences is similar to that between sequences that are merely part of the same gene family. More dramatically, homolog sequences can diverge “beyond recognition,” such that the similarity between two orthologs is not higher than the similarity between sequences that are not part of the same gene family and automatic procedures for the recognition of homology fail. A recent survey of genes in *Drosophila* shows that one-third of the cDNAs code for very fast evolving genes, for which the frequency of amino acid substituting mutations is only a 2-fold lower than that of silent mutations, leading to a situation where homologous proteins are barely recognizable after 8,000 years of evolution (16).

**Orthology in multidomain proteins.** Orthology in multidomain proteins is often more difficult to detect than orthology in single protein domains. A first step toward modularity, the presence in both genomes of neighboring sequences that are also orthologs of each other, then is considered orthologs of each other. As shown below, there is little conservation of the order of genes in genomes in evolution at a time when divergence of their orthologous genes reaches a level of 50% amino acid identity (see Fig. 3). Hence the potential for using synteny for identifying orthologs is limited mainly to genomes that have speciated only relatively recently. A second type of auxiliary information that can be used is the comparison of genes with those of a third genome. If two genes from different genomes have the highest level of identity both to each other and to a single gene from a third genome, then this is a strong indication that they are orthologs (see ref. 15 for a large-scale implementation of this idea). However for a large fraction of genes identifying orthologs by relative sequence identity is hampered by a variety of evolutionary processes. We describe these in the following sections.

**Sequence divergence.** At large evolutionary distances, e.g., between Archaea and Bacteria, sequence similarities may be eroded to such an extent that the distance between orthologous sequences is similar to that between sequences that are merely part of the same gene family. More dramatically, homolog sequences can diverge “beyond recognition,” such that the similarity between two orthologs is not higher than the similarity between sequences that are not part of the same gene family and automatic procedures for the recognition of homology fail. A recent survey of genes in *Drosophila* shows that one-third of the cDNAs code for very fast evolving genes, for which the frequency of amino acid substituting mutations is only a 2-fold lower than that of silent mutations, leading to a situation where homologous proteins are barely recognizable after 8,000 years of evolution (16).

**Orthology in multidomain proteins.** Orthology in multidomain proteins is often more difficult to detect than orthology in single protein domains. A first step toward modularity, the presence in both genomes of neighboring sequences that are also orthologs of each other, then is considered orthologs of each other. As shown below, there is little conservation of the order of genes in genomes in evolution at a time when divergence of their orthologous genes reaches a level of 50% amino acid identity (see Fig. 3). Hence the potential for using synteny for identifying orthologs is limited mainly to genomes that have speciated only relatively recently. A second type of auxiliary information that can be used is the comparison of genes with those of a third genome. If two genes from different genomes have the highest level of identity both to each other and to a single gene from a third genome, then this is a strong indication that they are orthologs (see ref. 15 for a large-scale implementation of this idea). However for a large fraction of genes identifying orthologs by relative sequence identity is hampered by a variety of evolutionary processes. We describe these in the following sections.

**Sequence divergence.** At large evolutionary distances, e.g., between Archaea and Bacteria, sequence similarities may be eroded to such an extent that the distance between orthologous sequences is similar to that between sequences that are merely part of the same gene family. More dramatically, homolog sequences can diverge “beyond recognition,” such that the similarity between two orthologs is not higher than the similarity between sequences that are not part of the same gene family and automatic procedures for the recognition of homology fail. A recent survey of genes in *Drosophila* shows that one-third of the cDNAs code for very fast evolving genes, for which the frequency of amino acid substituting mutations is only a 2-fold lower than that of silent mutations, leading to a situation where homologous proteins are barely recognizable after 8,000 years of evolution (16).
as one always keeps in mind the caveats described above and as long as the methods for determining orthology are well defined.

Timing Genome Divergence. To compare the rates at which various properties of genomes change, a central reference for the divergence between genomes is required. Measurement of the divergence times between the three “domains” (Archaea, Bacteria, and Eukarya) on the basis of protein dissimilarities recently has gained considerable attention and has been the subject of some controversy (see ref. 23 and references therein). The estimates of the date of the last common ancestor vary from 2 billion (24) to 3–4 billion years ago (23). The major assumptions in estimating divergence times from distances between protein sequences are: (i) The proteins are of vertical descent; i.e., they have not been horizontally transferred into the genome following the speciation of the species compared; and (ii) the proteins act as a molecular clock, having rates of amino acid substitutions that do not vary over time and between the lineages. Here we use proteins to scale divergence between and within the Archaea and the Bacteria. It is not our intention to estimate absolute divergence times, rather it is to compare the different relative rates at which genomes evolve. Thus we translate the protein dissimilarities between the species into amino acid substitutions per position per gene, using an equation derived by Grishin (25), which corrects for variations in substitution rates for both amino acids and sites: 

\[ q = \ln(1 + 2d)/2d \]

where \( q \) is the fraction of identical amino acids between the proteins and \( d \) is the number of amino acid substitutions per site. Grishin’s equation recently was used by Doolittle et al. (23) and gives reasonable estimates for the divergence between Bacteria and Archaea. Stringent criteria were used to select a set of genes that had orthologs in all of the nine genomes compared: (i) Each gene had the highest level of identity to at least five of the other genes (relative to other genes in those five genomes, see our minimal definition of orthology above); and (ii) there were no conflicting hits, from each genome only one protein was selected. The resulting set of 34 proteins is surprisingly small. It contains 17 ribosomal proteins, five tRNA synthetases, two signal recognition particles, two proteins with unknown function, and eight metabolic enzymes. Interestingly, the set consists almost exclusively of proteins that interact with RNA or synthesize RNA. In estimating divergence times of the genomes of Archaea and Bacteria it could be useful to check whether the protein similarities follow the phylogenetic tree (23) given the previously recognized ancient horizontal transfer of metabolic enzymes from Bacteria to Archaea (26), and more recent occasions of horizontal gene transfer (Fig. 1). However, because Archaeal genomes are chimeric, they were treated as

before the branching of A. fulgidus and M. thermoautotrophicum, supports this hypothesis. The only inconsistency is the fact that in the clustering of the kinases, AF1483 and slr0473 are slightly more similar to each other than either is to MTH444. (B) Domain architecture of slr0473, MTH444, and AF1483. The genes slr0473 and AF1483 are multidomain proteins, carrying GAF (43) domains and PAS (44, 45) motifs at their N terminus. The PAC motif (44, 45) could be detected only in AF1483. The GAF domain and PAC and PAC motifs are absent in MTH444, and have been replaced by three transmembrane regions (see ref. 11). All three genes possess a histidine kinase domain (HisKc) at their C terminus; 3’ to the slr0473 and MTH444 genes are the regulatory response genes slr0474 and MTH445. The distances between the reading frames are short: 15 nucleotides in Synechocystis and the reading frames overlap in M. thermoautotrophicum. In A. fulgidus the spatial association between these genes is absent. The absence of the GAF and PAS domains in MTH444 might have caused different selective constraints in MTH444 than in slr0473 and AF1493, and thus increased its rate of evolution, thereby reducing its similarity to its A. fulgidus and Synechocystis orthologs at a relatively high rate. The GAF, PAC, and PAS domains were predicted by using the SMART system (ref. 46; http://www.bork.embl-heidelberg.de/Modules/sinput.shtml).
such by obtaining a central reference for the distance between genomes by averaging over the proteins’ distances, irrespective of their phylogenetic trees. As Grishin’s equation tends to overestimate the number of amino acid substitutions per position for low levels of identities between genes (27), the median of the estimates of the number of amino acid substitutions was used in preference to the mean. The results are used in the following sections.

Comparing Genomes as “Bags of Genes”

Shared Orthologous Genes. The decrease of the number of shared orthologs in time. A straightforward comparison between genomes simply considers genes, and not the correlation between genes; i.e., a genome is regarded as a “bag of genes.” Taking this a step further, we measure how the number of shared orthologs between two genomes decreases with their divergence time (Fig. 2). The results show that the fraction of shared orthologous sequences decreases rapidly in evolution, faster than the level of pairwise identity between the shared orthologs. Although the fraction of shared orthologs between Archaea and Bacteria is less than among the Bacteria, the most dramatic reduction in the fraction of shared orthologs takes place on shorter time scales within the Bacteria and Archaea, when protein identity levels between genomes are still above 50%.

Non-tree-like aspects of the evolution of gene content. Even over large evolutionary distances such as those between Archaea and Bacteria different pairs of genomes share different orthologs. For example, M. genitalium shares different orthologs with M. jannaschii than with M. thermautotrophicum (see legend to Fig. 2). This demonstrates a nontree-like aspect of the evolution of the gene content of genomes: phylogenetically closely related species do not share orthologous genes that either of them shares with a phylogenetically distant species.

Differential Genome Analysis. Pairwise genome comparison. Instead of focusing on genomes’ similarities one can focus on their dissimilarities; i.e., “differential genome analysis” (28). Such analysis can be particularly revealing if the genomes are closely related but have different phenotypes, in which case one can identify the genetic basis for their differences. For example, of the genes in the pathogen H. influenzae that do not have a homolog in the relatively benign E. coli, a large fraction, 60% are (potentially) involved in H. influenzae’s pathogenesis (28). These genes encode proteins that are located on the surface of the cell or are involved in the production of toxins, or are virulence factors, or are homologous to proteins present only in pathogenic species. By contrast, of the proteins in H. influenzae that do have an ortholog in E. coli only an estimated 12% can be considered host interaction factors.

Multiple genome comparison. Differential genome analysis can be extended to multiple genomes. One then can analyze the correlation between shared gene content and shared phenotypic features of the species compared. This is demonstrated in a comparison of the two pathogens H. influenzae and H. pylori with E. coli. H. influenzae and H. pylori share 17 orthologs that do not have a homolog in E. coli. Of these, a large fraction (12) are related to pathogenicity (unpublished data). Differential genome analysis also can be used to select genes responsible for other differences in phenotypes, e.g., metabolism. The main requirement is that the genomes are sufficiently close in evolution that the identification of orthologs is reliable and that the differences in genome content reflect mainly the phenotypic feature that one is interested in.

Measuring Correlations Between Genes

Conservation of the Spatial Association of Genes. Quantification of the differentiation of gene order. Synteny, the con-

![Figure 2](image-url)
interaction indicates recent horizontal gene transfer events. Distant species of genes for proteins that do not show physical interaction are transferred together in horizontal gene transfer. Synteny between phylogenetically distant species of genes (see also ref. 31). As mentioned previously, there is experimental evidence for direct physical interaction that function well together will tend to increase, to prevent the separation of a coadapted pair of alleles by recombination.

Gene order and operons. Given the widely accepted concept of the operon, it is perhaps surprising that there is so little conservation of gene order. Why the gene order that is conserved only concerns proteins that show physical interaction might be explained by Fisher's model of gene clustering (35). Fisher argued that the linkage between genes of proteins that function well together will tend to increase, to prevent the separation of a coadapted pair of alleles by recombination.

It is clear that operons do not only exist of genes for proteins that show physical interaction (reviewed in ref. 36). However, what is conserved of operons over large time scales seems indeed to concur with Fisher's hypothesis. A theory that explains the rearrangement of operons has to include an explanation for the existence of operons. The overall rearrangement of operons does not support any theory that is based on functional relationships of the proteins coded by the genes in the operon, unless one specifically can show that functional relationships of the genes change over the time scales on which we observe the rearrangement of operons. The recently proposed theory of "selfish operons" proposes that operons exist because they increase the probability that genes that function together are transferred together in horizontal gene transfer (36). This model was based on the observation that operon structure is conserved between E. coli and Salmonella typhimurium. The model applies only to "nonessential" genes, genes that are relatively dispensable, which can be lost and then reintroduced into the genome through horizontal operon transfer. It, for example, does not apply to the ribosomal genes that are strongly clustered, are essential, and for which we have no evidence for horizontal gene transfer. It does, however, apply to pathogenicity islands and pathogenicity islets, clusters of genes that play a role in pathogenicity, and do indeed show evidence for horizontal gene transfer (37).

Regulatory Elements. With the determination of orthologous genes and conservation of gene order one can begin to determine whether intergenic regions are conserved. The degree of conservation of intergenic regions is remarkably low and is diverging much faster than the gene order (Y. Diaz-Lazcoz, M.A.H. and P.B., unpublished results). The pattern in Fig. 4 can be regarded as an exception, demonstrating that at least in some cases gene regulation is preserved. At the 5' end of the ribosomal genes rpl11 and rpl1 in E. coli lies an RNA secondary structure potentially involved in the regulation of expression of the rpl11 operon (38). The structure is conserved...
in all Bacterial genomes analyzed in this paper, with the notable exception of *H. pylori*.

**Co-Occurrence of Genes.** Some genomes are more organized than others. If neighboring genes tend to function together in one genome, as they do in the case of operons, then they should both occur in another genome, even if they are not neighbors or part of the same operon. We show (Fig. 5A) that this is indeed the case. If gene A has a neighboring gene B, then if the ortholog of B (B’) occurs in another genome the probability that the ortholog of A (A’) occurs in the other genome is increased (compare Fig. 2). In other words, orthologs shared between two genomes tend to be clustered in at least one of the genomes. Part of the results of Fig. 5A are caused by genes that occur as neighbors in both of the genomes compared. The analysis was repeated to only include genes that are separated in one genome (X), but neighbors in another genome (Y). The fraction of genes that are neighbors in Y was compared with the expected fraction, given a model of random shuffling of genes (see Fig. 5B for methods). Results show that genes from a genome Y that have an ortholog in genome X tend to cluster in Y. The trend is present in all genomes except *M. genitalium*, and is particularly pronounced in the genomes of *E. coli* and *B. subtilis*. This surprising result suggests that most genomes are organized, yet some genomes are more organized than others. We assume that the genes that occur in one genome and are neighbors in another genome are in some way or another related in function. One explanation for the high degree of clustering in *E. coli* and *B. subtilis* is that they consist to a large fraction of recent horizontal gene transfers, which could increase the prevalence of polycistronic operons in their genome.

**Co-occurrence of genes and the conservation of pathways.** Instead of analyzing spatial association of orthologs, one can analyze whether orthologs show “genome association”: i.e., they either occur together in a genome or are both absent from a genome. Such an analysis could, in principle, be used to reconstruct which genes are functionally related. The fact that orthologs that both occur in two genomes have a relative high probability of spatial association in one of the genomes (Fig. 5A), even if they are separated in the other genome (Fig. 5B), in itself points to the usefulness of this idea. By analogy to approaches using the covariance of the nucleotide content of positions in RNA (39) to predict which positions interact with each other, one can use the covariance in the occurrence of proteins to create a model of which proteins depend for their function on each other. Such information could be used to reconstruct metabolic pathways or signaling pathways. The important assumption is that the structure of the pathway was constant throughout evolution. Nonorthologous gene displacement, where a gene assumes the functions of another in a pathway suggests that pathways are more conserved than the presence of orthologous genes. Our observation of the co-occurrence of the genes dnaJ and dnaK in a small set of orthologs that are shared by *M. genitalium* and *M. thermotolerificum*, but not by *M. jannaschii* (see legend Fig. 2), dnaK shows that the correlation of functionally related genes is present in phylogenetically distant species.

The existence of associated genes and the conservation of this association are important parameters in determining the degree of epistasis of genome evolution and determine the shape of the “adaptive landscape” (40) in which genome evolution operates. For an analysis of covariation in the occurrence of genes to be statistically meaningful more genomes then the nine that were analyzed here are required. Furthermore one needs to correct for the “baseline” probability that a gene from one genome has an ortholog in another genome, which depends on phylogenetic distance between the genomes (Fig. 2).

### Comparing Rates of Genome Evolution

We have studied several indicators of genome evolution and followed their conservation over time (Fig. 6). The resulting calibration curves do quantify not only the divergence of these indicators, but also have practical value as they show what information can be extracted from new microbial genomes.


---

**Fig. 6.** Relative rates of genome evolution. The curves were fitted from the fraction of shared orthologs (Fig. 2) and the conservation of the order of genes (Fig. 3), the curve that shows the relationship between protein identity and the number of amino acid substitutions per position according to Grishin’s equation (Fig. 2), was added for comparison. Intergenic regions are even less conserved than the order of genes. Nonorthologous gene displacement indicates that metabolism is more conserved than the fraction of shared orthologous genes.

---

We are very grateful to Chris Ponting, Berend Snel, Yolande Diaz-Lazcoz, Thomas Dandekar, and Joerg Schultz for providing data and useful discussions. The work was supported by the Bundesministerium für Bildung, Wissenschaft, Forschung and Technologie (Germany) and Deutsche Forschungsgemeinschaft.