Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 5849–5856, May 1998
Colloquium Paper

This paper was presented at the colloquium "Computational Biomolecular Science," organized by Russell Doolittle, J. Andrew McCammon, and Peter G. Wolynes, held September 11–13, 1997, sponsored by the National Academy of Sciences at the Arnold and Mabel Beckman Center in Irvine CA.

Measuring genome evolution

(ortholog/synteny/computer analysis/horizontal gene transfer)

MARTIJN A. HUYNEN* 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 thermoautotrophicum* (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

*To whom reprint requests should be addressed at: European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany. e-mail: huynen@embl-heidelberg.de.

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 orthologous.

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).

Nonorthologous gene displacement. A second event problematic to ortholog identification is nonorthologous gene displacement. This occurs when two nonorthologous genes that are unrelated or only remotely related perform the same function in two organisms (17). This occurs relatively frequently: a comparison of *M. genitalium* to *H. influenzae* revealed 12 clear-cut cases (17). As a consequence orthologs may not be detectable (or are classified as paralogs) in another organism even when the corresponding function is retained.

Gene duplication, gene loss, and horizontal gene transfer. A third process that restricts the identification of orthologous genes is that of gene loss in combination with gene duplication. If two genomes lose different paralogs of an ancestral gene that was duplicated before the speciation event, the remaining genes have highest sequence identity even though they are not orthologs (18). One may test for such an event by checking whether the protein similarity falls into an expected range. This is done implicitly by including (presumably orthologous) sequences from other species in the phylogeny and checking whether the gene tree is in accordance with the species tree (18, 19). Inconsistencies between the species tree and the gene tree can indicate nonorthologous relationships between genes. However, they also can be caused by horizontal gene transfer, in which case the genes still could be orthologs. In general, the identification of orthologous sequences, horizontal gene transfer, and ancient gene duplications cannot be distinguished. Besides the construction of phylogenetic trees an additional strategy for finding horizontal gene transfer is the comparison of nucleotide frequencies within a genome. Recently transferred genes often display nucleotide frequencies that deviate significantly from the rest of the genome (20, 21). A conservative estimate of the amount of genes that recently have been transferred to *E. coli*, based on nucleotide frequencies and dinucleotide frequencies in genomes is 10%−15% of the *E. coli* genome (Phil Green, personal communication; ref. 21). A third strategy for finding horizontal gene transfer is synteny. Because gene order is rarely conserved in evolution, the presence in two distant evolutionary branches of the same order of genes, combined with the absence of this gene order in other more closely related branches, can point to horizontal gene transfer. This strategy has been used successfully to find the example of horizontal gene transfer described in Fig. 1.

Orthology in multidomain proteins. In multidomain proteins two levels of orthology can be distinguished: one is at the level of single domains, a second at the level of the whole protein. This may lead to situations where nonorthologous proteins possess orthologous domains. Modularity of genes in the sense that modules can have different positions, but the same function, in various proteins, is not well documented in Bacteria and Archaea. A first step toward modularity, the presence of "gene fusion" or "gene splitting," however, does occur regularly. Comparative analysis of the genomes *H. influenzae* and *E. coli* showed 10 (24) clear-cut cases of genes that were separate in *E. coli* (*H. influenzae*), but that were part of a single gene in *H. influenzae* (*E. coli*) (unpublished data).

A much more complicated scenario, for which many of the factors described above (multidomain proteins, synteny, and horizontal gene transfer) are involved, is shown in Fig. 1. In general, a combination of the various evolutionary processes described above leads to a situation where, although orthology was defined originally as a one-to-one relationship between proteins, it must be considered a many-to-many relationship.

From homologs to orthologs. The advent of powerful, easy-to-use tools, such as PSI-BLAST (22), to find homologous sequences is likely to shift the emphasis in sequence analysis from predicting homology to predicting orthology. It is clear that, at present, there is not a single, simple, and perfect solution to the question of orthology. Orthology is methodologically defined, that is, dependent on what is asked of the genomes that are compared, different methods to find orthologous genes are being used. We use a minimal definition when we are interested only in the number of orthologs shared between genomes at various phylogenetic distances. Orthologs then are defined in the following manner: (*i*) They have the highest level of pairwise identity when compared with the identities of either gene to all other genes in the other's genome; (*ii*) the pairwise identity is significant (E, the expected fraction of false positives, is smaller than 0.01), and (*iii*) the similarity extends to at least 60% of one of the genes. The region of similarity is not required to cover the majority of both genes to include the possibility of gene fusion and gene splitting. In more detailed comparisons between a small number of genomes, auxiliary information was used to determine orthology, such as the order of genes and the comparison to genes from a third genome (see legend to Fig. 1).

Given all of these complications in the finding of orthologs and the oversimplified view of evolution that the term suggests, one could conclude that it is better not to use it at all, or only in those cases where one does not have conflicting information from various sources about the phylogeny of the genes. One also can argue that it is exactly these cases where there are conflicts in the information about orthology from different sources that evolution shows some of its most interesting aspects. Orthology is an important refinement over homology in describing the phylogenetic relations between genes, as long

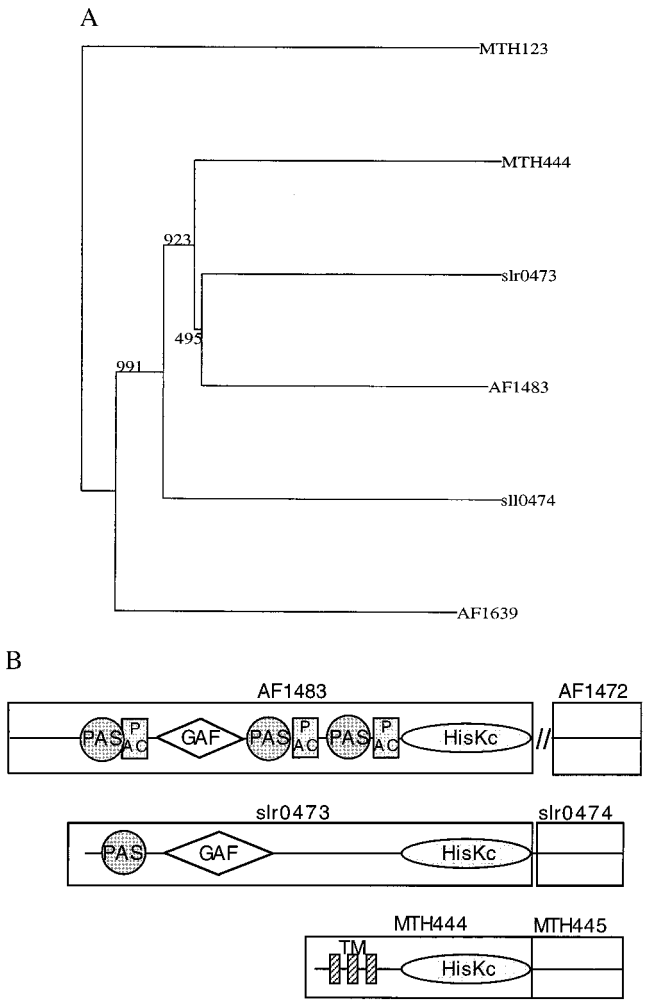

FIG. 1. An example of complexities in assigning orthology to multidomain proteins. The *M. thermoautotrophicum* genes MTH444 (a sensory transduction histidine kinase) and MTH445 (a sensory transduction regulatory protein) are orthologs of the *Synechocystis* sequences slr0473 (phytochrome; ref. 41) and slr0474, respectively (the gene nomenclature is from the GenBank files of complete genomes, the first letters of gene names generally represent the initials of the genomes). The arguments for orthology are: (*i*) The genes have a 34.8% and a 40.2% identity to each other, which is significantly higher than either of them has to other sequences in the other's genome. (*ii*) They are neighboring genes in both genomes. (*iii*) Both MTH444 and slr0473 have the highest level of identity to a single sequence from a third species *Archeoglobus fulgidus* (42), AF1483, the same is true for MTH445 and slr0474 with respect to AF1472. Interestingly, the level of identity of the *Synechocystis* sequences slr0473 and slr0474 is significantly higher to the *M. thermoautotrophicum* and *A. fulgidus* sequences than it is to any of the sequences in the Bacteria, including sequences in *Synechocystis* itself. The reverse is even more dramatic: MTH445, AF1472, and MTH444, AF1483 are more identical, not only to their *Synechocystis* orthologs, but also to 27 respectively 28 other sequences in *Synechocystis* than they are to sequences in their own genomes. These 27 (28) sequences are paralogs of slr0473 (slr0474). The similarity between MTH444 and AF1483 is slightly lower than that between AF1483 and slr0473, whereas the similarity between AF1472 and MTH444 is significantly higher than that of either of them to slr0473. Neighbor-joining clusterings of the histidine kinase orthologs together with their most similar sequences from the three genomes (*A*) illustrates the most likely evolutionary scenario: a horizontal transfer of the genes in the branch that has led to *Synechocystis*, to the branch leading to *M. thermoautotrophicum* and *A. fulgidus*. Given the relative similarities of the proteins, this event occurred after a major amplification of the histidine kinase family in *Synechocystis* and not long before the split of the branches that led to *M. thermoautotrophicum* and *A. fulgidus*. The fact that none of the proteins have a detectable homolog in *M. jannaschii*, which branched off in the Archaea not long 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, AF1483, and MTH444. 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 PAS 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 AF1483, 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).

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

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. thermoautotrophicum* (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 conservation of the order of genes, has been extensively studied already. Although some conservation of the order of genes in genomes has been reported (29, 30), the emphasis has been on the the drastic rearrangement of gene order in evolution (31–33). The evolution of the spatial organization of the genome is being studied for three reasons: (*i*) To calibrate the rate at which it evolves. (*ii*) To study the genome organization of the last common ancestor (34). Shared gene order between the Archaea and the Bacteria is assumed to date back to their last common ancestor, with the exception of horizontal gene transfer (Fig. 1). (*iii*) To estimate the time scale at which gene regulation changes during evolution. The spatial association of genes is related to their regulation, e.g., in the case of operons.

The conservation of gene order was related to genome divergence time (Fig. 3). The results show a drastic rearrange-

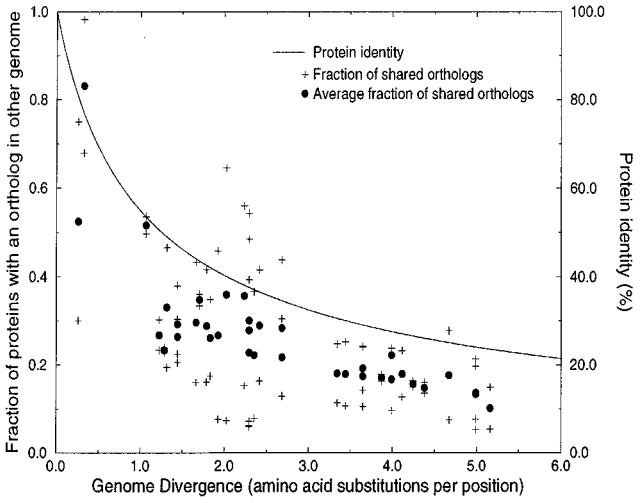

FIG. 2. The relationship between genome similarity, measured as the fraction of shared orthologs, and time, measured as the number of amino acid substitutions per protein per position in a set of 34 orthologs. + shows the fraction of sequences in a genome A that has an ortholog in another genome B, and vice versa. This measure is asymmetric, a relatively small genome like *H. influenzae* is more similar to a large one like *E. coli* than *E. coli* is similar to *H. influenzae*. ● shows the average of the two asymmetric similarities. Here we use a minimal definition of orthology: sequences that between two genomes have the highest, significant ($E < 0.01$) level of pairwise identity, that covers at least 60% of one of the proteins are regarded as orthologs. Sequences were compared with the Smith–Waterman algorithm (47), using a parallel Bioccellerator computer. The relationship between sequence identity and the number of amino acid substitutions per position as calculated with Grishin's equation (25) is given for comparison. If one assumes that the divergence time between the Archaea and Bacteria is 3.5 billion years (23), the unit of one amino acid substitution corresponds to about 875 million years. In this estimate of divergence time the Mycoplasmas and *H. pylori* are not included, because they have a relatively high rate of evolution. The highest six divergence times correspond to the comparisons of the Mycoplasmas and *H. pylori* with the Archaea. As is clear from the figure, the fraction of shared orthologs between genomes decreases more rapidly in evolution than does the protein identity. Note that the base level of shared orthologs at which the figure saturates consists only partly of a set of sequences that are shared by all the genomes compared. For example, there are 15 orthologous pairs shared between *M. genitalium* and *M. thermoautotrophicum* of which none of the genes has a homolog at the $E < 0.01$ level in *M. jannaschii*. Of this set, the ones with the highest level of protein identity are: DnaK and DnaJ (MG305 and MG019), heat shock proteins with 51% and 50% identity, respectively to their *M. thermoautotrophicum* ortholog, deoxyribose-phosphate aldolase (MG050) with 40% identity, a pyrophosphatase (MG351) with 40.5% identity, and a transcriptional regulator (MG448) with 45% identity. Genes that are shared by *M. genitalium* and *M. jannaschii* but that are absent in *M. thermoautotrophicum*, include proteins from the glycolysis like pyruvate kinase (MG216) with 29.1% identity and glucose-6-phosphate isomerase (MG111) with 27% protein identity.

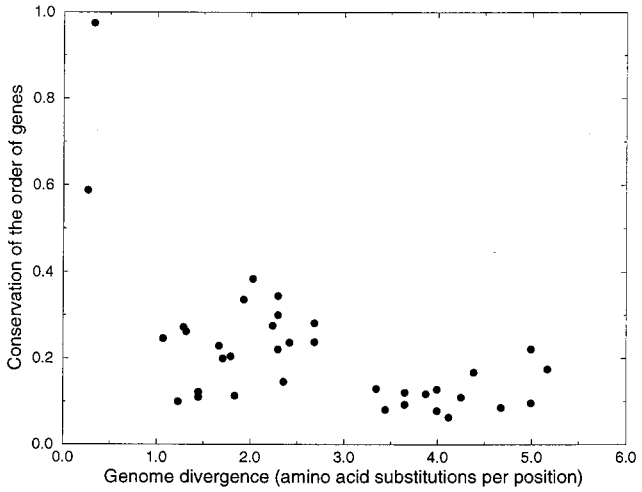

FIG. 3. Conservation of the order of genes within the genome. Shown are the number of genes that are orthologs in both genomes, and that have at least one neighboring gene that is the same ortholog in both genomes, divided by the total number of shared orthologs between the genomes. The *x* axis shows the divergence of the genomes measured in amino acid substitutions per position. The figure clearly indicates the rapid differentiation of gene order in evolution. Gene order between genomes is less conserved than the fraction of shared orthologs (compare with Fig. 2).

ment of genomes within the first time unit, during which protein identity levels remain above 50%, after which a saturation level is reached. Notice that the order of orthologous genes is less preserved than their presence (compare with Fig. 2). At the divergence time at which the saturation level is reached, the genes that are still paired are in general subunits of proteins, ribosomal proteins or proteins involved in ABC transport. A detailed examination (T. Dandekar, M.A.H., and P.B., unpublished data) of all conserved pairs of proteins in three Gram-negative bacteria (*E. coli*, *H. influenzae*, and *H. pylori*) and in three Archaea (*M. thermoautotrophicum*, *M. jannaschii*, and *A. fulgidus*) has shown that, for nearly all cases, there is experimental evidence for direct physical interaction between these proteins (see also ref. 31). As mentioned previously, this observation has implications for the study of horizontal gene transfer. Synteny between phylogenetically distant species of genes for proteins that do not show physical interaction indicates recent horizontal gene transfer events.

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 co-adapted 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

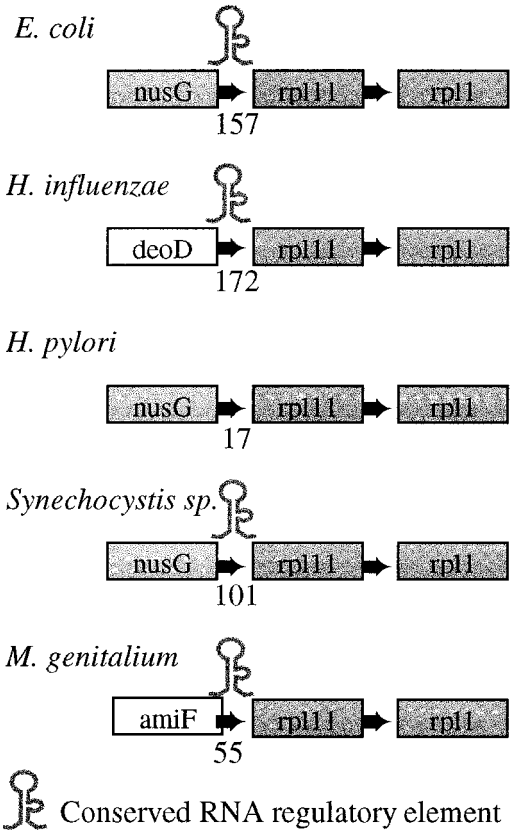

FIG. 4. Conservation of an RNA secondary structure at the 5′ end of rpl11 operon in Bacterial genomes. The order of the ribosomal protein genes *rpl11* and *rpl1* is conserved in all of the Bacteria analyzed. The gene *nusG* is a transcription antitermination factor, Amif is an oligopeptide transport ATP-binding protein, and *deoD* codes for a purine-nucleoside phosphorylase. The number between the first and second gene indicates the length of the intergenic region. Surprisingly, the secondary structure is absent from *H. pylori*, even though it shares the presence of *nusG* 5′ of *rpl11* with *E. coli*, whereas *H. influenzae* lacks *NusG* at this position. Notice furthermore that the element has been deleted in *H. pylori* rather than lost because of point mutations, as there is no space left between *nusG* and *rpl11* in *H. pylori*. The element is also present in *M. pneumoniae*, but is absent from the Archaea. The element is part of the 5′ leader of the L11 mRNA sequence and is likely to function in the autoregulation of the *rpl11* operon (ref. 38 and Y. Diaz-Lazcoz, M.A.H. and P.B., unpublished data).

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. 5*A*) 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. 5*A* 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. 5*B* 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 results 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 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. 5*A*), even if they are separated in the other genome (Fig. 5*B*), in itself points to the usefulness of this idea. By analogy to approaches using the covariation of the nucleotide content of positions in RNA (39) to predict which positions interact with each other, one can use the covariation 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. thermoautotrophicum*, 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 epistatis 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).

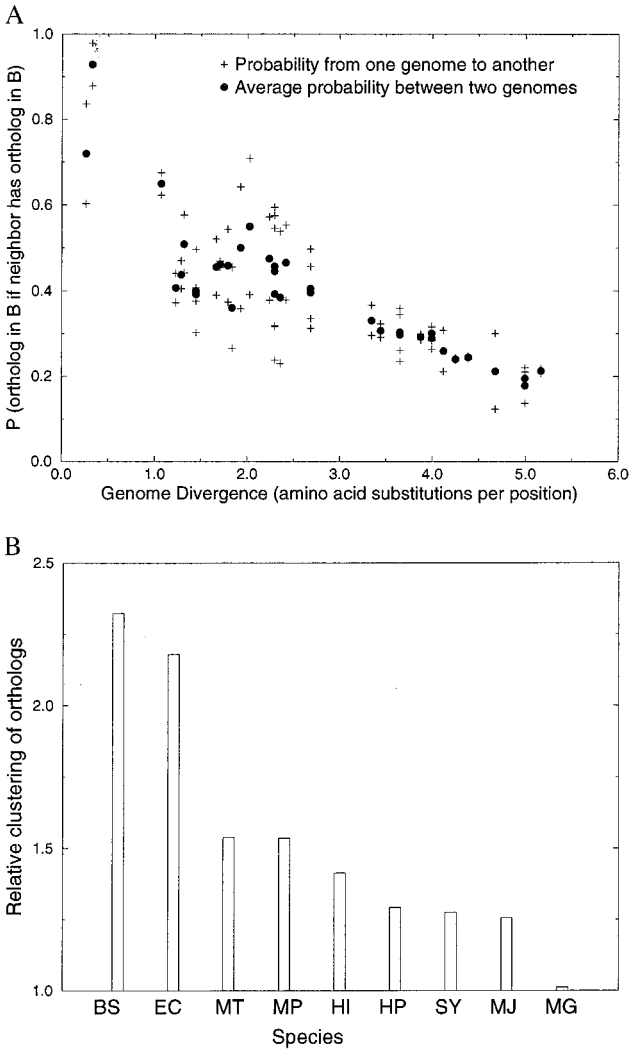

FIG. 5. (*A*) The probability that a gene in genome A has an ortholog in another genome B if a neighboring gene in A has an ortholog in genome B. The probabilities clearly increase, as compared with the average probability of having an ortholog in another genome (compare Fig. 2). (*B*) The relative degree of clustering of genes in one genome (A) that have an ortholog in another genome (B). The analysis includes only genes that are clustered ("neighbors") in genome A, but not in B (and vice versa). Shown is the ratio of the number of genes in A that have an ortholog in B and have at least one neighboring gene that also has an ortholog in B, divided by the expected number. The expected number of genes that are neighbors in a genome, given a random distribution, is calculated as follows: Given X genes that are randomly distributed over a genome with Y loci, the probability that a gene from X has no neighboring genes from X (it lies isolated) is the probability that it has no left-neighbor from X nor a right-neighbor from X: $P_0 = [(Y - X)/(Y - 1)]^* [(Y - X - 1)/(Y - 2)]$. The expected number of genes from X with at least one neighbor from X: $P_{1,2} = 1 - P_0$. The fraction of genes in genome A with at least one neighbor that also has an ortholog in genome B is thus divided by $P_{1,2}$ to get to the relative clustering of the genes in genome A. The relative clustering is averaged over the genome comparisons of one genome versus the eight other genomes. The names of the species have been abbreviated to the first letters of their genus and species name. All genomes, except *M. genitalium* show a more than expected clustering of genes. Given its small size, *M. genitalium* has relatively little room to cluster the genes that have an ortholog in another genome above the expected level of clustering: i.e., most of the genes that have an ortholog in another genome are expected to be neighbors in *M. genitalium*. The correlation with genome size is not perfect however. For example, *Synechocystis*, which has a relatively large genome, shows relatively little genome organization.

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

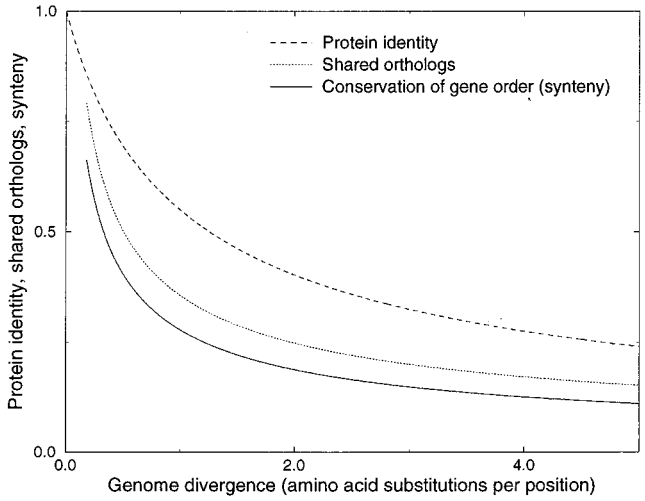

FIG. 6. Relative rates of genome evolution. The curves were fitted from the fraction of shared orthologs (Fig. 2) and the conservation of the order of genes (Fig. 3), the curve that shows the relationship between protein identity and the number of amino acid substitutions per position according to Grishin's 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.

given their phylogenetic position. The calibration curves shall require refinement when more data become available but they already provide levels of expectation, deviations from which are of potential interest (e.g., synteny of genes in distant species that cannot be found in other species is an indicator of horizontal gene transfer; Fig. 1). In particular, more relatively closely related genomes that have protein identity levels higher than 50% shall be essential to provide more precise estimates of the rates at which genome organization and gene regulation evolve. The calibration curves also should influence the analysis strategy, e.g., if a closely related genome is available, orthologs are relatively easy to discriminate from other members of multigene families. By analogy to profile search techniques, it is helpful to include not too closely related but also not too divergent species into the first round of the analysis, where the closeness of the relationship depends on the features one wants to identify. For example, to study the evolution of gene regulation one needs to compare more closely related species than to study the evolution of gene order. To study the evolution of gene content, one needs to compare even less related species, whereas the study of the evolution of metabolism requires the comparison of the most distantly related species.

Current analysis of genomes is driven by the prediction of functional features at the molecular and cellular level; it is based on the presence and absence of certain genes in the context of phenotypic expectations. Expectations about horizontal gene transfers and the loss, the acquisition or displacement of entire pathways (the entire metabolism in the case of the Archaea) and the study of the correlations of gene occurrence will enable us to identify functional cascades in greater detail. Identification of weak regulatory signals in the genomes requires a sensitive comparative analysis. The puzzling evolution of nonconserved but ever-present operons is only one indication that many genetic and evolutionary mechanisms are yet to be detected and quantified.

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.

1. Blattner, F. E., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & Mayhew, G. F. (1997) *Science* **277**, 1453–1462.
2. Karlin, S., Mrazek, J. & Campbell, A. (1997) *J. Bacteriol.* **179**, 3899–3913.
3. Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleishmann, R. D., Venter, J. C. & Moxon, E. R. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 11121–11125.
4. Huynen, M. A. & van Nimwegen, E. (1998) *Mol. Biol. Evol.*, in press.
5. Gelfand, M. S. & Koonin, E. V. (1997) *Nucleic Acids Res.* **25**, 2430–2439.
6. Fleishmann, R., Adams, M., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & Merrick, J. M. (1995) *Science* **269**, 496–512.
7. Fraser, C. M., White, O., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., Ketchum, K. A., Dodson, R. & Hickey, E. K. (1995) *Science* **270**, 397–403.
8. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M. & Sasamoto, S. (1996) *DNA Res.* **3**, 109–136.
9. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A. & Gocayne, J. D. (1996) *Science* **273**, 1058–1072.
10. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. & Herrmann, R. (1996) *Nucleic Acids Res.* **24**, 4420–4449.
11. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R. & Gilbert, K. (1997) *J. Bacteriol.* **17**, 7135–7155.
12. Tomb, J.-F., White, O., Kervalage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A. (1997) *Nature (London)* **388**, 539–547.
13. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A. & Borchert, S. (1997) *Nature (London)* **390**, 249–256.
14. Fitch, W. M. (1970) *Syst. Zool.* **19**, 99–110.
15. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278**, 631–637.
16. Schmid, K. & Tautz, D. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 9746–9750.
17. Koonin, E. V., Mushegian, A. R. & Bork, P. (1996) *Trends Genet.* **12**, 334–336.
18 Page, R. D. M. (1994) *Syst. Biol.* **43**, 58–77.
19. Yuan, Y. P., Eulenstein, O., Vingron, M. & Bork, P. (1998) *Bioinformatics*, in press.
20. Medigue, C., Rouxel, Y., Vigier, P., Henaut, A. & Danchin, A. (1991) *J. Mol. Biol.* **222**, 851–856.
21. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44**, 383–397.
22. Althschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
23. Feng, D. F., Cho, G. & Doolittle, R. F. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 13028–13033.
24. Doolittle, R. F., Seng, D. F., Tsang, S., Cho, G. & Little, E. (1996) *Science* **271**, 470–477.
25. Grishin, N. V. (1995) *J. Mol. Evol* **41**, 675–679.
26. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997) *Mol. Microbiol.* **25**, 619–637.
27. Feng, D.-F. & Doolittle, R. F. (1997) *J. Mol. Evol.* **44**, 361–370.
28. Huynen, M., Diaz-Lazcoz, Y. & Bork, P. (1997) *Trends Genet.* **13**, 389–390.
29. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. & Koonin, E. V. (1996) *Curr. Biol.* **6**, 279–291.
30. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997) *J. Mol. Evol.* **44**, 66–73.
31. Mushegian, A. R. & Koonin, E. V. (1996) *Trends Genet.* **12**, 289–290.
32. Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. (1997) *J. Mol. Evol.* **44**, 57–64.
33. Kolsto, A. B. (1997) *Mol. Microbiol.* **24**, 241–248.
34. Siefert, J. L., Martijn, K. A., Abdi, F., Widger, W. R. & Fox, G. E. (1997) *J. Mol. Evol.* **45**, 467–472.
35. Fisher, R. A. (1930) *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford).
36. Lawrence, J. G. & Roth, J. R. (1996) *Genetics* **143**, 1843–1860.
37. Barinaga, M. (1996) *Science* **272**, 1261–1263.

38. Branlant, C., Krol, A., Machatt, A. & Ebel, J. P. (1981) *Nucleic Acids Res.* **9**, 293–307.
39. Gutell, R. R., Power, A., Hertz, G., Putz, E. & Stormo, G. (1993) *Nucleic Acids Res.* **20**, 5785–5795.
40. Wright, S. (1932) in *Proceedings of the Sixth International Congress on Genetics*, ed. Jones, D. F. (Brooklyn Botanical Garden, New York), Vol. 1, pp. 356–366.
41. Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. (1997) *Science* **277**, 1505–1508.
42. Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K. & Peterson, J. D. (1997) *Nature (London)* **390**, 364–370.
43. Aravind, L. & Ponting, C. P. (1997) *Trends Biochem. Sci.* **22**, 458–45.
44. Zhulin, I. B., Taylor, B. L. & Dixon, R. (1997) *Trends Biochem. Sci.* **22**, 331–333.
45. Ponting, C. P. & Avarind, L. (1997) *Curr. Biol.* **7**, R674–R677.
46. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864.
47. Smith, T. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.